BMC Genomics

RESEARCH Open Access

Molecular pathway identification using biological network-regularized logistic models

Wen Zhang[1,2†], Ying-wooi Wan[1,2,3†], Genevera I Allen[2,4], Kaifang Pang[1,2], Matthew L Anderson[3,6], Zhandong Liu[1,2,5,6*]

From The International Conference on Intelligent Biology and Medicine (ICIBM 2013)
Nashville, TN, USA. 11-13 August 2013

Abstract

Background: Selecting genes and pathways indicative of disease is a central problem in computational biology. This problem is especially challenging when parsing multi-dimensional genomic data. A number of tools, such as L_1-norm based regularization and its extensions elastic net and fused lasso, have been introduced to deal with this challenge. However, these approaches tend to ignore the vast amount of *a priori* biological network information curated in the literature.

Results: We propose the use of graph Laplacian regularized logistic regression to integrate biological networks into disease classification and pathway association problems. Simulation studies demonstrate that the performance of the proposed algorithm is superior to elastic net and lasso analyses. Utility of this algorithm is also validated by its ability to reliably differentiate breast cancer subtypes using a large breast cancer dataset recently generated by the Cancer Genome Atlas (TCGA) consortium. Many of the protein-protein interaction modules identified by our approach are further supported by evidence published in the literature. Source code of the proposed algorithm is freely available at http://www.github.com/zhandong/Logit-Lapnet.

Conclusion: Logistic regression with graph Laplacian regularization is an effective algorithm for identifying key pathways and modules associated with disease subtypes. With the rapid expansion of our knowledge of biological regulatory networks, this approach will become more accurate and increasingly useful for mining transcriptomic, epi-genomic, and other types of genome wide association studies.

Introduction

Technologies for high throughput genetic profiling have revolutionized the study of human development and disease. Expression profiles spanning the entire human genome not only allow investigators to better understand disease subtypes [1], but also define new categories associated with sensitivity to pharmacologic treatment [2,3] and other clinical outcomes [4,5]. A central problem in these genomic studies is to construct an accurate predictive model and delineate specific genes or pathways driving a phenotype. Logistic regression is widely used for classification [6-9]. The number of genes, however, in high-throughput studies are often much larger than the number of specimens in a given study. This limitation causes instability in the algorithms used to select driver genes and poor performance of predictive models. The lasso algorithm for logistic regression was introduced to address these problems and perform feature selection through L_1-norm regularization [10]. However, the number of variables selected by the lasso is bounded by the number of observations in an experiment. Furthermore, correlated variables are rarely selected as part of the predictive model at the same time. Various extensions of the Lasso algorithm, such as elastic net, pelora, grouped lasso and fused lasso introduce grouping or smoothness regularization terms to address these limitations [11-14]. Of these, the fused

* Correspondence: zhandonl@bcm.edu
† Contributed equally
[1]Department of Pediatrics-Neurology, Baylor College of Medicine, Houston, TX, USA
Full list of author information is available at the end of the article

lasso and elastic net have been successfully applied to a largest number of gene expression and genome wide association studies [15-18]. Both fused Lasso and elastic net allow correlated genes or neighboring genes on a chromosome to be selected into a predictive model together. However, these algorithms tend to ignore functional interactions between individual gene products documented in the scientific literature. Integration of network information in the gene marker identification has shown to outperform methods without network information by Chuang et al [19]. Specifically, their networked-based scoring and greedy search algorithm identified more robust gene markers with better prediction accuracy on metastasis status of breast cancer patients in two cohorts. In recent years, vast amounts of data detailing biologic networks has been organized into searchable databases. For example, BioGRID documents protein and genetic interactions from more than 39,991 publications [20]. The KEGG pathway database similarly curates molecular interactions and relational networks representing systemic functions at the level of both the cell and organism [21]. To incorporate biological network information into regression models, a network-constrained regularization algorithm has been previously proposed for use with linear regression [22]. Use of this network-constrained algorithm has been shown to out-perform both lasso and elastic net analyses executed independently of biologic input.

Classification algorithms integrating network structure information have been proposed in other settings. For example, a network constrained support vector machine has been proposed to analyze functional magnetic resonance imaging data [23] and cancer microarray data [24-26]. Network based prior has also been demonstrated to improve variable selection accuracy under the Bayesian inference framework [27,28].

Here, we propose a graphical Laplacian network regularized logistic regression method following the framework established by Li et al. [22]. We hypothesize that the integration of biological networks, such as Protein-Protein interactions, will improve prediction accuracy and variable selection in logistic models. To validate use of the proposed algorithm, we studied its theoretical properties and compared its performance to L_1-norm regularized logistic regression and elastic net logistic regression on both simulated and real biologic data. Lastly, we demonstrate the utility of the proposed algorithm by using it to differentiate breast cancer subtypes and delineate biologic network modules associated with the triple negative breast cancer (TNBC) subtypes.

Materials and methods

Graph laplacian regularized logistic regression model

Suppose that the data set has n observations with p genes. Let $\mathbf{y} = (y_1, ..., y_n)^T$ be the response with $y_i \in \{0, 1\}$ and $\mathbf{X} = [\mathbf{x_1}|, ..., |\mathbf{x_p}]$ be the matrix of biomarkers measured on n samples with $\mathbf{x_j} = (x_{1j}, ..., x_{nj})^T$ for $j = 1, ..., p$ genes. Without loss of generality, we can assume that each gene is standardized. The binary clinical variable, $\mathbf{y}$, can be predicted using the logistic model:

$$\text{Prob } (\mathbf{y} = 1|\mathrm{X}; \beta) = \frac{1}{1 + e^{-\mathrm{X}\beta}} \tag{1}$$

The parameter β can be estimated by maximizing the log likelihood function of the logistic model. However, it is well known that this estimation procedure performs poorly for both prediction purposes and variable selection when $p \gg n$. Although various sparse parameter estimation procedures have been introduced to address these problems, these approaches tend to generate disconnected biomarkers that are rarely interpretable.

To incorporate biological network information into the model estimation procedure, we adopted the network-constrained regularization framework proposed by Li et al [22]. Given a biological network $G = (V, E)$, where V is the set of genes that correspond to p predictors, E is the adjacency matrix, and $e_{uv} = 1$ if there exists an edge between u and v, otherwise $e_{uv} = 0$. The normalized graph Laplacian matrix L for G can be defined by

$$L = \mathrm{I} - D^{-\frac{1}{2}} E D^{-\frac{1}{2}} \tag{2}$$

where $D = \mathrm{diag}(E{\cdot}1_p)$ is a degree matrix with the diagonal elements equal to the degrees for each node in G. For any fixed non-negative regularization parameters λ and α, we define the logistic graph Laplacian net (Logit-Lapnet) criteria as:

$$\mathbb{L}(\lambda, \alpha, \beta) = \sum_{i=1}^{n} [-y_i X_i \beta + \ln(1 + e^{X_i \beta})] \; + \; \lambda\alpha|\beta|_1 + \lambda(1-\alpha)\langle\beta, \beta\rangle_L \tag{3}$$

Where

$$|\beta|_1 = \sum_{j=1}^{p} |\beta_j|,$$

$$\langle\beta, \beta\rangle_L = \beta^T L \beta = \sum_{e_{uv} \neq 0} \left(\frac{\beta_u}{d_u} - \frac{\beta_v}{d_u}\right)^2.$$

d_u, d_v are the degrees of nodes u and v respectively. In (3), X_i is the ith row of the matrix X. The first term in equation (3) is the negative log likelihood function of the logistic model. The second term is an L_1-norm penalization on β, which encourages sparsity on the coefficients. The last term is a generalized L_2-norm penalty using the graph Laplacian matrix, which encourages smoothness on coefficients of genes that are connected in the biological network. If we set α = 1, the Logit-Lapnet criteria is equal to the simple lasso logistic regression. When $L = \mathbf{I}$ which is identity matrix, the Logit-Lapnet criteria becomes the elastic net logistic regression.

The model estimation can be formulated into a convex optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \mathbb{L}(\lambda, \alpha, \beta). \tag{4}$$

To solve problem (4) we used CVX, a package for specifying and solving convex programs [29,30]. The convexity property of Logit-Lapnet guarantees an optimal solution using any convex optimization solver.

Theoretical properties of logit-lapnet

To study the behavior of the proposed method, we analyzed the theoretical properties of Logit-Lapnet.

Lemma 1 *If* $\mathbf{x}_i = \mathbf{x}_j$*, then* $\hat{\beta}_i = \hat{\beta}_j$*for any* $\alpha < 1$ *and* $\lambda > 0$.

The proof of this lemma is given in the additional file 1. This lemma states that if two predictors are identical, the coefficients of these two predictors will be the same. In the L_1-norm penalized logistic regression, only one predictor is selected.

Theorem 1 *Given data* (**y**, **X**) *and parameters* (λ, α)*, let* $\hat{\beta}(\lambda, \alpha)$ *be the Logit-Lapnet estimator for problem (4). If* $\hat{\beta}_i\hat{\beta}_j > 0$ *and* $e_{ij} = 1$*, define* $D_{\lambda,\alpha}(i,j) = \frac{1}{|y|_1}|\hat{\beta}_i(\lambda,\alpha) - \hat{\beta}_j(\lambda,\alpha)|$, *then*

$$D_{\lambda,\alpha}(i,j) \leq \frac{\sqrt{2(1-\rho)}}{2\lambda(1-\alpha)} \tag{5}$$

where ρ *is the sample correlation of* $\mathbf{x}_i$ *and* $\mathbf{x}_j$.

The upper bound in (5) provides a quantitative description for the grouping effect of Logit-Lapnet on the network structure. For two highly correlated genes ($\rho = 1$) that are connected in a biological network, the difference on the estimated coefficient is almost zero. As α goes to 1, the Lapnet becomes lasso logistic regression and the difference becomes unbounded.

Lemma 2 *The maximum value of* λ *in problem (4) with* $\hat{\beta} \neq 0$*satisfies*

$$\lambda_{\max} \leq \frac{1}{2\alpha}(2|X^T y|_\infty + |\sum_{i=1}^{n} X_i^T|_\infty) \tag{6}$$

The maximum value of λ is reciprocal to α. It is clear that when $\alpha = 1$, the Logit-Lapnet criteria becomes Lasso and a small penalization can produce an empty model. When $\alpha = 0$, the penalization becomes much larger to generate an empty model. In practice, Lemma 2 provides a search guidance on the regularization path. The proof of Lemma 2 is also given in the additional file 1.

Gene expression profiles and differential gene analysis

Gene expression data from breast cancer specimens profiled by TCGA consortium was used to test the proposed Logit-Lapnet method. Level III RNA-Seq data (Illumina HigSeq RNASeqV2 from UNC) from patients with invasive breast carcinoma was obtained from the TCGA data portal https://tcga-data.nci.nih.gov/tcga/ in September, 2012. These data profiled 20501 genes in 806 distinct breast cancer specimens. Normalized read counts were used for all analyses and were log-transformed prior to their use. Genes with normalized read counts less than 10 in more than 90% of patients were excluded from further analyses. The normalized read counts were log-transformed and standardized prior to applying the method.

Of the 806 breast cancers characterized by the TCGA, 261 patients had incomplete information on the three markers used to define TNBC and were excluded. Of the remaining cancers, 85 TNBC and 460 non-TNBC were further separated into training (n = 327; 51 TNBC, 276 non-TNBC) and test (n = 218; 34 TNBC, 184 non-TNBC) sets. A total of 4871 differentially expressed gene products (P ≤ 0.05, t-test; ≥ 1.5 fold-change between TNBC vs. non-TNBC on training) were merged with protein-protein interaction (PPI) networks. This provided us with 797 genes and PPI networks of 937 interactions available for analysis using our proposed algorithm.

Protein-protein interaction (PPI) network

To construct the graph Laplacian matrix for testing the Logit-Lapnet method on breast cancer data, networks of protein-protein interactions (PPI) identified by two-hybrid screening in Homo sapiens were obtained from The Biological General Repository for Interaction Datasets (BioGRID, version 3.2.98) http://www.thebiogrid.org. At the time it was accessed, BioGRID documented 21483 interactions between 7700 genes.

Results and discussion

Simulation studies

We initially used a benchmark simulation proposed by Li to explore the performance of the proposed Logit-Lapnet algorithm [22]. In brief, a network with 200 distinct transcription factors (TFs) was simulated. Each TF in this simulation regulated 10 genes with a total of 2,200 genes in the simulated network. The clinical variable y was assigned a binary value and was associated with the first four TFs and their target genes. In model I, we assumed that two of the TFs and their targets were positively associated with the clinical variable and the other two TFs and their targets were negatively associated with the clinical variable.

$$\beta = (5, \underbrace{\frac{5}{\sqrt{10}}, \ldots, \frac{5}{\sqrt{10}}}_{10}, -5, \underbrace{\frac{-5}{\sqrt{10}}, \ldots, \frac{-5}{\sqrt{10}}}_{10}, 3, \underbrace{\frac{3}{\sqrt{10}}, \ldots, \frac{3}{\sqrt{10}}}_{10}, -3, \underbrace{\frac{-3}{\sqrt{10}}, \ldots, \frac{-3}{\sqrt{10}}}_{10}, 0, \ldots, 0)^T$$

The clinical variable was defined as y = [Prob (y = 1| X; β) > ϱ] where ϱ ~ $U(0,1)$. Expression levels for the

200 TFs were then simulated using a standard normal distribution. Each TF and its target genes were jointly distributed as a bivariate normal with correlation of 0.7.

In model II, gene expression levels were simulated similarly to model I except that a TF could be both a transcriptional activator and repressor at the same time. The coefcient vector was defined as:

$$\beta = (5, \frac{-5}{\sqrt{10}}, \frac{-5}{\sqrt{10}}, \frac{-5}{\sqrt{10}}, \underbrace{\frac{5}{\sqrt{10}}, \ldots, \frac{5}{\sqrt{10}}}_{7}, -5, \frac{5}{\sqrt{10}}, \frac{5}{\sqrt{10}}, \frac{5}{\sqrt{10}}, \underbrace{\frac{-5}{\sqrt{10}}, \ldots, \frac{-5}{\sqrt{10}}}_{7}, 3, \frac{-3}{\sqrt{10}}, \frac{-3}{\sqrt{10}}, \frac{-3}{\sqrt{10}},$$
$$\underbrace{\frac{3}{\sqrt{10}}, \ldots, \frac{3}{\sqrt{10}}}_{7}, -3, \frac{3}{\sqrt{10}}, \frac{3}{\sqrt{10}}, \frac{3}{\sqrt{10}}, \underbrace{\frac{-3}{\sqrt{10}}, \ldots, \frac{-3}{\sqrt{10}}}_{7}, 0, \ldots, 0)^T$$

Model III was similar to model I except that we decreased the association of the target genes on the clinical variable and made the model even sparser.

$$\beta = (5, \underbrace{\frac{5}{10}, \ldots, \frac{5}{10}}_{10}, -5, \underbrace{-\frac{5}{10}, \ldots, -\frac{5}{10}}_{10}, 3, \underbrace{\frac{3}{10}, \ldots, \frac{3}{10}}_{10}, -3, \underbrace{-\frac{3}{10}, \ldots, -\frac{3}{10}}_{10}, 0, \ldots, 0)^T$$

Model IV was similar to model II in allowing transcription factors to function as both activators and repressors. However, the clinical association of the target genes was decreased.

$$\beta = (5, -\frac{5}{10}, -\frac{5}{10}, -\frac{5}{10}, \underbrace{\frac{5}{10}, \ldots, \frac{5}{10}}_{7}, -5, \frac{5}{10}, \frac{5}{10}, \frac{5}{10}, \underbrace{-\frac{5}{10}, \ldots, -\frac{5}{10}}_{7}, 3, -\frac{3}{10}, -\frac{3}{10}, -\frac{3}{10},$$
$$\underbrace{\frac{3}{10}, \ldots, \frac{3}{10}}_{7}, -3, \frac{3}{10}, \frac{3}{10}, \frac{3}{10}, \underbrace{-\frac{3}{10}, \ldots, -\frac{3}{10}}_{7}, 0, \ldots, 0)^T$$

For each model, we simulated both a training data set as well as an independent test data set of 100 samples. A 10-fold cross-validation procedure was applied to the training data set to identify the optimal tuning parameter. Genes with non-zero coefficient in the estimated model were found to be associated with the clinical variable. The sensitivity and specificity of our variable selection performance was defined as the following:

$$\begin{aligned} &\text{True Negative (TN)} := |\bar{\beta}. * \bar{\hat{\beta}}|_0, \qquad \text{False Positive (FP)} : = |\bar{\beta}. * \hat{\beta}|_0 \\ &\text{False Negative (FN)} := |\beta. * \bar{\hat{\beta}}|_0, \qquad \text{True Positive (TP)} : = |\beta. * \hat{\beta}|_0 \\ &\text{Sensitivity} : = \frac{\text{TP}}{\text{TP + FN}}, \quad \text{Specificity} : = \frac{\text{TN}}{\text{TN + FP}} \end{aligned} \tag{7}$$

where the $|\cdot|_0$ counts the number of non-zero elements in a vector, $\bar{\beta}$ is the logical not operator on a vector and.* is the element-wise product.

We repeated the experiment 50 times. Results are summarized for each model in Table 1. We also computed the Bernoulli error loss on the test data set. For all four models, we compared the performance of the proposed algorithm to both the L_1 penalized logistic regression and the elastic net algorithm. Our method resulted in much higher sensitivity in identifying associated genes with the same specificity compared to the other two algorithms (Table 1). Our method also gave much smaller MSE compared to the Lasso and elastic net logistic regression. We computed the Receiver Operator Curve on the whole regularization path for each of the algorithm. In all four models, the proposed algorithm demonstrated much higher precision compared to Lasso and elastic net logistic regression (Figure 1).

Biomarker identification using logit-lapnet in human breast cancers

Breast cancers are clinically categorized according to the expression of several gene products, including the estrogen receptor (ESR1), progesterone receptor (PGR) and human epidermal growth factor receptor 2 (ERBB2). These biomarkers are routinely used not only to define prognosis but also determine treatment. [31,32]. Of the breast cancer subtypes defined by these biomarkers, triple negative breast cancers (TNBC) lacking expression of ESR1, PAR and ERBB2(her2) are the most clinically aggressive. TNBC demonstrate high rates of disease progression and recurrence [33]. Outcomes for this breast cancer subtype are generally poor, largely due to the fact that treatment options for women with TNBC are limited. However, a subgroup of TNBC are highly sensitive to conventional chemotherapy [34].

We chose to next validate use of our method by testing its ability to identify the key biomarkers used to define breast cancer subtypes. To accomplish this goal, we used patterns of gene expression in a large set of

Table 1 Results of simulations.

	Sensitivity			Specificity			Bernoulli Error Loss		
#	Lasso	Elastic	Lapnet	Lasso	Elastic	Lapnet	Lasso	Elastic	Lapnet
1	0.3146	0.4995	**0.7583**	0.9744	**0.9945**	0.9832	19.24	17.2	**13.84**
	(0.059)	(0.069)	(0.071)	(0.0028)	(0.002)	(0.005)	(0.578)	(0.537)	(0.493)
2	0.1852	0.4386	**0.6577**	**0.9982**	0.9936	0.9847	19.22	19.62	**16.8**
	(0.042)	(0.076)	(0.079)	(0.001)	(0.003)	(0.005)	(0.702)	(0.671)	(0.571)
3	0.2614	0.5714	**0.8832**	0.9887	**0.9893**	0.9537	15.14	16.44	**14.38**
	(0.045)	(0.068)	(0.066)	(0.0024)	(0.0026)	(0.006)	(0.582)	(0.585)	(0.495)
4	0.2314	0.6755	**0.8645**	**0.9927**	0.9583	0.9492	17.38	18.86	**16.1**
	(0.043)	(0.0712)	(0.069)	(0.0019)	(0.0045)	(0.009)	(0.511)	(0.572)	(0.570)

The results for simulations. Sensitivity, specificity and PMSEs are based on 50 simulations. The standard errors are given in parentheses.

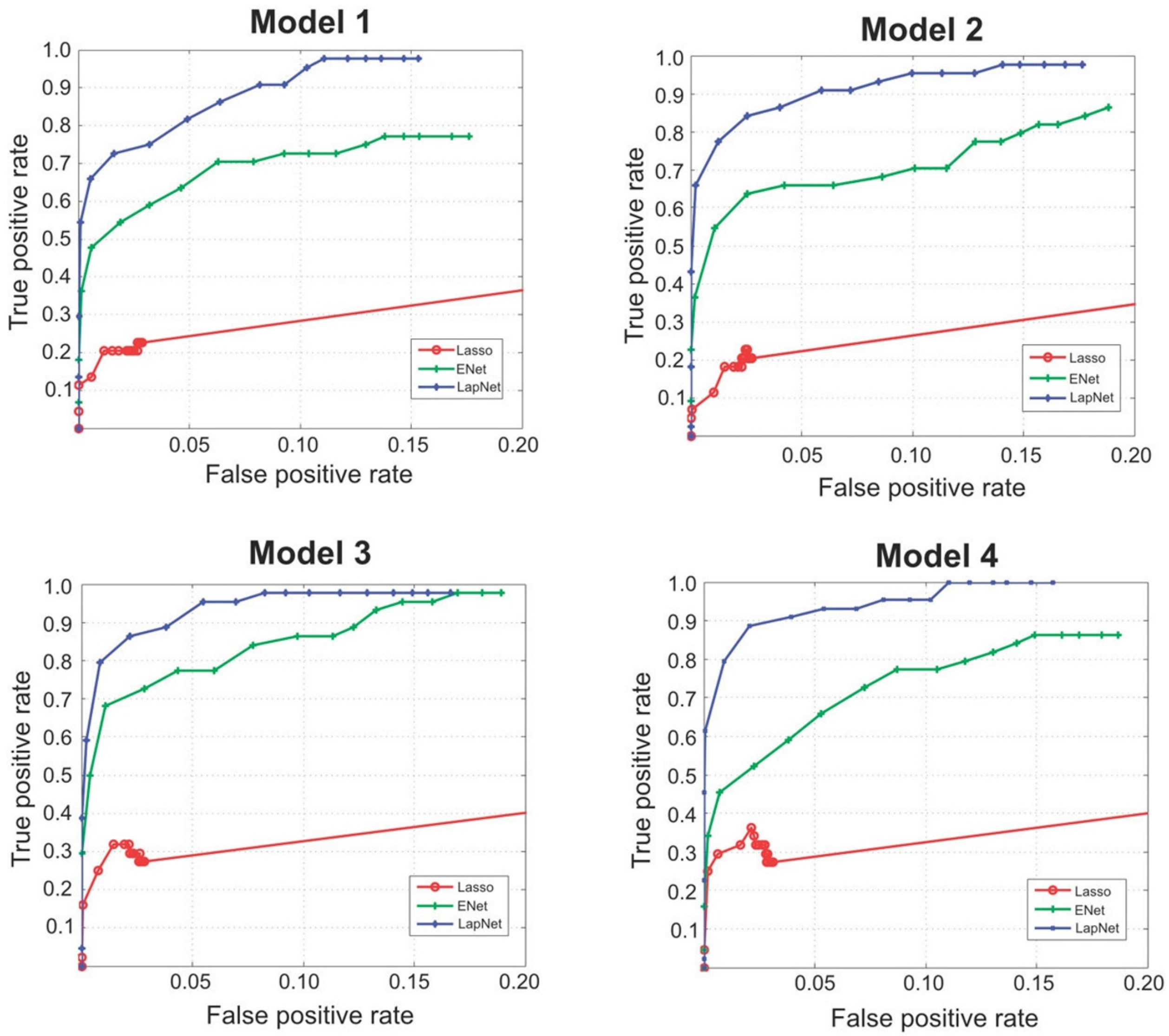

Figure 1 Receiver operator curves (ROCs) on the regularization path. ROCs were computed on the regularization path for Logit-Lapnet, enet, and lasso for all four models (A)-(D). Logit-Lapnet has higher true positive rate and lower false positive rate compared to the other two approaches.

breast cancer specimens recently profiled by the Cancer Genome Atlas consortium. The tuning parameter for each method was selected through 10-fold cross-validation using the training data. We observed the same classification accuracy in predicting TNBC tumors (95%) when lasso, elastic net and Logit-Lapnet were applied to the testing data. The similar efficacy of these 3 algorithms is not surprising since TNBC tumors are dramatically different from the non-TNBCs in terms of their gene expression profiles.

However, there were a number of significant differences observed between the results obtained with the 3 algorithms tested. For example, use of Logit-Lapnet selected 262 genes, of which > 63% (166 genes) interacted with one another (Figure 2A). In comparison, use of the Lasso algorithm selected only 24 genes, most of which (20 genes) were isolated and were not predicted to interact. Elastic net selected nearly half of the input genes (393 genes) of which 59% (230 genes) were interconnected (Additional File 2). Furthermore, neither Lasso nor elastic net identified the progesterone receptor as a key discriminator for TNBC, despite the fact that this gene product is routinely used to clinically categorize breast cancers. Only Logit-Lapnet successfully identified each of the three markers used to define TNBC subtype: ESR1, PGR and ERBB2. These results suggest that Logit-Lapnet is more accurate than either Lasso or elastic net for identifying biomarkers from

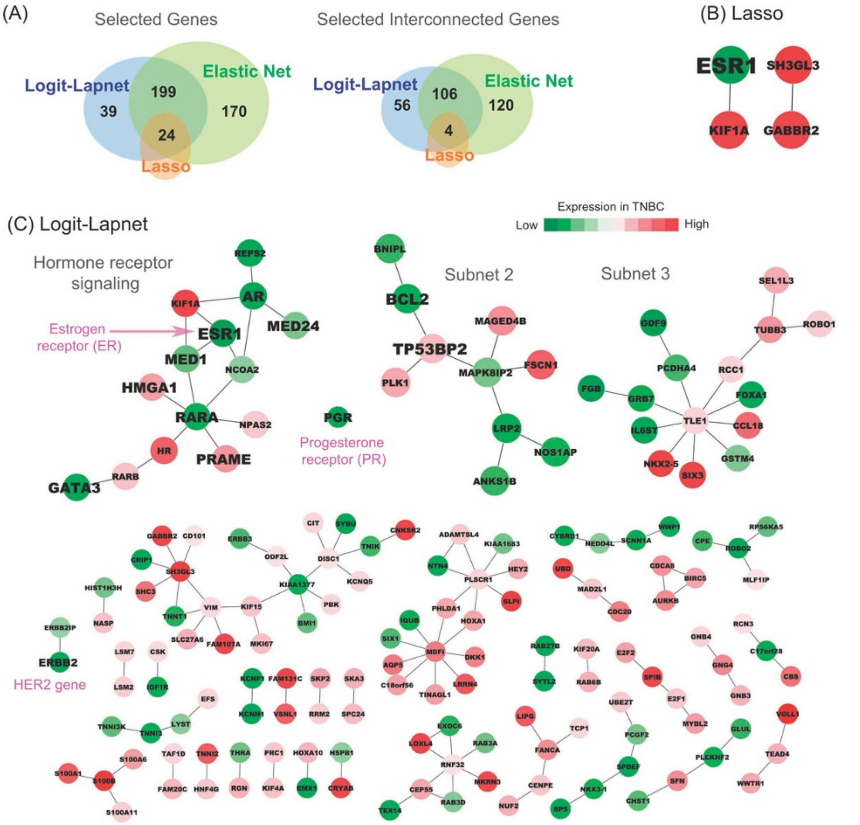

Figure 2 Application of the algorithm to identify TNBC-associated genes. Genes and subnetworks of PPI associated with TNBC using TCGA breast cancer data and BioGRID PPI. Comparison of the selected genes from our proposed algorithm with those from lasso and elastic net (A). Genes and their respective subnetworks of PPI recovered by lasso (B) and our proposed Laplacian net algorithm (C). In the networks, genes having association to breast cancer reported in the literature are labeled with larger font.

large multidimensional genomic datasets such as those generated by the TCGA.

Use of logit-lapnet for inferring novel biologic relationships

In addition to identifying each of the 3 key biomarkers used to categorize breast cancer subtypes, we found that our method identified multiple subnetworks of gene expression in the breast cancer specimens profiled by TCGA. These subnetworks potentially reflect relationships with clinical or biologic significance. For example, one of subnetworks we identified includes multiple genes (AR, ESR1, MED1, MED24, RARA, PRAME, and HMGA1) involved in steroid hormone signaling. As demonstrated in Figure 2, Logit-Lapnet found that this cluster is closely connected to a known breast cancer gene, GATA3. This connection suggests a functional relationship. Evidence to support this relationship could not be found by directly searching the database of protein-protein interactions used to construct our algorithm. However, at least one report published since our initial analyses has now shown that GATA3 mediates genomic ESR1-binding upstream of FOXA1 [35]. This confirms that the integration of genomic and PPI data by our method has the capacity to identify new and otherwise unanticipated relationships with biologic significance.

Enhanced network specificity provided by logit-lapnet

Another advantageous feature of Logit-Lapnet is the potential functional specificity of the subnetworks delineated by our algorithm. Subnetworks identified by Logit-Lapnet allow investigators to more readily focus on key genes for subsequent downstream functional analyses. For example, the subnetwork connecting AR, ESR1 and RARA identified by our algorithm in TNBC represents a cluster of genes involved in hormone receptor signaling (Figure 2). The subnetwork 2 identified by our method contains the genes PLK1, BCL2, BNIPL, and tumor suppressor p53 binding protein TP53BP2. PLK1, BNIPL and BCL2 are well-known oncogenes and part of TP53 pathway [36]. The expression of BCL2 has been proposed as a prognostic marker for breast cancer patients. This subnetwork is also interesting, as it predicts a relationships between gene products that have been previously shown to impact the G1-S (TP53BP) and G2-M cell cycle checkpoints (PLK1). Dysregulation of both G1-S and G2-M are key hallmarks of human cancer as defined by Weinberg and others. Furthermore, TP53 dysfunction has been previously shown to lead to the overexpression of PLK1 and other gene products important for driving cells with genomic instability through the cell cycle. Thus, the ability of Logit-Lapnet to detect this relationship underscores its capacity to detect key events in breast and other human cancers. These results convey a clear message that interactions between genes in clusters might be biologically relevant.

In contrast, alternative algorithms such as elastic net identify fewer subnetworks containing larger number of genes. This means that subnetworks identified by Logit-Lapnet allow investigators to more readily focus on key targets for subsequent downstream functional validation. Although network modularization methods can be used to further dissemble networks defined by elastic net into smaller modules, these smaller modules may not reflect direct interactions between their individual components. This is because elastic net defines groups of co-expressed genes rather than networks of functionally interacting gene products. Alternatively, use of Logit-Lapnet can be considered to place gene correlations in the context of biologic function. Thus, its use can be reasonably anticipated to define relationships that are more likely to be biologically and clinically relevant.

In summary, the proposed algorithm identified set of genes associated with breast tumors. In addition, integration of PPI in the algorithm enables us to recover the genes not only in association to the breast cancer subtypes but also their interacting partners which are also breast cancer related. Most importantly, from the comparison with elastic net and lasso, our method selected a reasonable size of genes and is the only algorithm capable of identified all three marker genes in defining TNBC.

Conclusion

We have developed a graph Laplacian regularized logistic regression model for selecting the genes or network modules associated with clinical variables from gene expression profiles. Through simulation studies, we have demonstrated that our approach is much more sensitive for identifying clinically relevant genes and network modules. We have presented a case study on network module identification for triple negative breast cancer sub-type using mRNA-seq data. Our results indicate that Logit-Lapnet is a superior algorithm compared to Lasso and elastic net in disease gene and network module selection. Further work to biologically validate our predicted modules is needed to gain a more complete picture of the regulatory process in the TNBC sub-type. Beyond genomics, there are many potential applications of Logit-Lapnet to utilize the rich information provided by network structure, such as metabolomics and proeteomics studies. In conclusion, our work developing the network structure regularized logistic regression model has many implications and has provided a new tool for research in genomic studies.

Additional material

Additional file 1: Proofs of lemma and theorem. This file includes the mathematics proofs on lemma 1, 2 and theorem 1.

Additional file 2: Genes identified by elastic net. This file includes a figure of genes and their respective subnetworks of PPI identified by elastic net. Genes with larger font indicates its association to breast cancer reported in the literature.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Genevera Allen and Zhandong Liu developed the model and wrote the paper. Wen Zhang, Ying-Wooi Wan and Kaifang Pang collected the real application data and conducted the application. Matthew Anderson provided biologic interpretation of networks and assisted with writing the paper. Zhandong Liu designed the simulation and real applications. Wen Zhang and Zhandong Liu realize and analyze the simulations.

Acknowledgements

This work is supported in part through the Collaborative Advances in Biomedical Computing Seed Funding Program at the Ken Kennedy Institute for Information Technology at Rice University supported by the John and Ann Doerr Fund for Computational Biomedicine and through the Center for Computational and Integrative Biomedical Research Seed Funding Program at Baylor College of Medicine. GA is also partially supported by NSF DMS-1209017. ZD is supported by the Houston Bioinformatics Endowment and NSF DMS-1263932.

Declarations

The publication fee is supported by the Houston Endowment.
This article has been published as part of *BMC Genomics* Volume 14 Supplement 8, 2013: Selected articles from the International Conference on Intelligent Biology and Medicine (ICIBM 2013): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/14/S8.

Authors' details

[1]Department of Pediatrics-Neurology, Baylor College of Medicine, Houston, TX, USA. [2]Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, TX, USA. [3]Department of Obstetrics and Gynecology, Baylor College of Medicine, Houston, TX, USA. [4]Department of Statistics and Electrical Engineering, Rice University, Houston, TX, USA. [5]Computational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, TX, USA. [6]Dan L.DuncanCancerCenter, BaylorCollegeofMedicine,Houston,TX,USA.

Published: 9 December 2013

References

1. Golub TR, Slonim DK, Tamayo P, Huard C: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *science* 1999, **286(5439)**:531-537.
2. Natsoulis G, El Ghaoui L, Lanckriet GRG, Tolley AM, Leroy F, Dunlea S, Eynon BP, Pearson CI, Tugendreich S, Jarnagin K: **Classification of a large**

microarray data set: algorithm comparison and analysis of drug signatures. *Genome research* 2005, **15(5)**:724-736.

3. Holleman A, Cheok MH, den Boer ML: **Gene-Expression Patterns in Drug-Resistant Acute Lymphoblastic Leukemia Cells and Response to Treatment.** *New England Journal of Medicine* 2004, **351(6)**:533-542, [PMID: 15295046].
4. Veer Lv, Dai H, Van De Vijver MJ, He YD: **Gene expression profiling predicts clinical outcome of breast cancer.** *nature* 2002, **415(6871)**:530-536.
5. Wang Y, Klijn J, Zhang Y, Sieuwerts AM, Look MP: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *Lancet* 2005, **365**:671-679.
6. Zhu J, Hastie T: **Classification of gene microarrays by penalized logistic regression.** *Biostatistics* 2004, **5(3)**:427-443.
7. Shen L, Tan EC: **Dimension reduction-based penalized logistic regression for cancer classification using microarray data.** *IEEE/ACM Trans Comput Biol Bioinform* 2005, **2**:166-175.
8. Liao JG, Chin KV: **Logistic regression for disease classification using microarray data: model selection in a large p and small n case.** *Bioinformatics* 2007, **23(15)**:1945-1951.
9. Bootkrajang J, Kabán A: **Classification of mislabelled microarrays using robust sparse logistic regression.** *Bioinformatics* 2013, **29(7)**:870-877.
10. Tibshirani R: **Regression Shrinkage and Selection Via the Lasso.** *Journal of the Royal Statistical Society, Series B* 1994, **58**:267-288.
11. Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005, **67(2)**:301-320.
12. Dettling M, Bühlmann P: **Finding predictive gene groups from microarray data.** *Special Issue on Multivariate Methods in Genomic Data Analysis* 2004, **90**:106-131.
13. Yuan M, Lin Y: **Model selection and estimation in regression with grouped variables.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2006, **68**:49-67.
14. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K: **Sparsity and smoothness via the fused lasso.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005, **67**:91-108.
15. Wu TT, Chen YF, Hastie T, Sobel E, Lange K: **Genome-wide association analysis by lasso penalized logistic regression.** *Bioinformatics* 2009, **25(6)**:714-721.
16. Kim S, Xing EP: **Statistical estimation of correlated genome associations to a quantitative trait network.** *PLoS genetics* 2009, **5(8)**:e1000587.
17. Tian Z, Zhang H, Kuang R: **Sparse Group Selection on Fused Lasso Components for Identifying Group-Specific DNA Copy Number Variations.** *ICDM'12* 2012, 665-674.
18. Liu J, Huang J, Ma S, Wang K: **Incorporating group correlations in genome-wide association studies using smoothed group Lasso.** *Biostatistics* 2013, **14(2)**:205-219.
19. Chuang HY, Lee E, Liu YT, Lee D, Ideker T: **Network-based classification of breast cancer metastasis.** *Mol Syst Biol* 2007, **3**:140.
20. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Research* 2006, **34(suppl 1)**:D535-D539.
21. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic Acids Research* 2012, **40(D1)**:D109-D114.
22. Li C, Li H: **Network-constrained regularization and variable selection for analysis of genomic data.** *Bioinformatics* 2008, **24(9)**:1175-1182.
23. Grosenick L, Klingenberg B, Katovich K, Knutson B, Taylor JE: **Interpretable whole-brain prediction analysis with GraphNet.** *NeuroImage* 2013, **72**:304-321.
24. Chen L, Xuan J, Riggins R, Clarke R, Wang Y: **Identifying cancer biomarkers by networkconstrained support vector machines.** *BMC systems biology* 2011, **5**:161.
25. Zhu Y, Shen X, Pan W: **Network-based support vector machine for classification of microarray samples.** *BMC Bioinformatics* 2009, **10(Suppl 1)**: S21.
26. Pan W, Xie B, Shen X: **Incorporating predictor network in penalized regression with application to microarray data.** *Biometrics* 2010, **66(2)**:474-484.
27. Stingo FC, Chen YA, Tadesse MG, Vannucci M: **Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes.** *The Annals of Applied Statistics* 2011, **5(3)**:1978-2002.
28. Hill S, Neve R, Bayani N, Kuo WL, Ziyad S, Spellman P, Gray J, Mukherjee S: **Integrating biological knowledge into variable selection: an empirical Bayes approach with an application in cancer biology.** *BMC bioinformatics* 2012, **13**:94.
29. Grant M, Boyd S: **Graph implementations for nonsmooth convex programs.** In *Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences.* Springer-Verlag Limited;Blondel V, Boyd S, Kimura H 2008:95-110[http://www.stanford.edu/~boyd/papers/pdf/graph_dcp.pdf].
30. CVX Research I: **CVX: Matlab Software for Disciplined Convex Programming, version 2.0.** 2012 [http://cvxr.com/cvx].
31. Siegel R, Naishadham D, Jemal A: **Cancer statistics, 2013.** *CA: A Cancer Journal for Clinicians* 2013, **63**:11-30.
32. The TCGA: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490(7418)**:61-70.
33. Stevens KN, Vachon CM, Couch FJ: **Genetic susceptibility to triple-negative breast cancer.** *Cancer Res* 2013, **73(7)**:2025-2030.
34. Foulkes WD, Smith IE, Reis-Filho JS: **Triple-negative breast cancer.** *N Engl J Med* 2010, **363(20)**:1938-1948.
35. Theodorou V, Stark R, Menon S, Carroll JS: **GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility.** *Genome Research* 2013, **23**:12-22.
36. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4(3)**:177-183.

doi:10.1186/1471-2164-14-S8-S7
Cite this article as: Zhang *et al.*: Molecular pathway identification using biological network-regularized logistic models. *BMC Genomics* 2013 **14** (Suppl 8):S7.