Software and Application Note

Nikos Vlassis and Enrico Glaab*

GenePEN: analysis of network activity alterations in complex diseases via the pairwise elastic net

Abstract: Complex diseases are often characterized by coordinated expression alterations of genes and proteins which are grouped together in a molecular network. Identifying such interconnected and jointly altered gene/protein groups from functional omics data and a given molecular interaction network is a key challenge in bioinformatics. We describe GenePEN, a penalized logistic regression approach for sample classification via convex optimization, using a newly designed Pairwise Elastic Net penalty that favors the selection of discriminative genes/proteins according to their connectedness in a molecular interaction graph. An efficient implementation of the method finds provably optimal solutions on high-dimensional omics data in a few seconds and is freely available at http://lcsb-portal.uni.lu/bioinformatics.

Keywords: machine learning; microarray analysis; network analysis.

DOI 10.1515/sagmb-2014-0045

1 Introduction

Genes and proteins belonging to the same region of a molecular network have a tendency to undergo coordinated expression alterations in complex diseases (Ideker and Sharan, 2008). We present GenePEN, an algorithm that identifies groups of genes/proteins forming connected network components as predictive features for disease-control sample classification tasks by using high-throughput expression data and an input graph encoding pairwise functional associations between corresponding biomolecules (e.g. protein-protein interactions). GenePEN is implemented in TFOCS, an open-source Matlab-based optimization language (Becker et al., 2011), and is designed for high efficiency (a few seconds runtime for a dataset with 100 samples and 10,000 features).

Cross-validation results on microarray gene expression data for both neurological and cancer diseases (Parkinson's disease, prostate and colorectal cancer) show marked improvements in terms of the network grouping of selected features as compared to related methods (measured by the size of the largest connected network component and compared against the related methods Elastic Net, the different variants of the Pairwise Elastic Net and Lasso). This network-grouping of selected features is not obtained at the expense of predictive power or detection rates for truly differential features, as shown by comparative evaluations on real-world and simulated data. The utility of the identified gene sub-networks for biological data

***Corresponding author: Enrico Glaab,** University of Luxembourg, Luxembourg Centre for Systems Biomedicine, 7, avenue des Hauts Fourneaux, Esch-sue-Alzette 4362, Luxembourg, Phone: +352 46 66 44 6186, e-mail: enrico.glaab@uni.lu
Nikos Vlassis: Adobe Research, Systems Technology Lab/Imagination Lab, 345 Park Avenue, San Jose, CA 95110, USA

interpretation is confirmed by an over-representation analysis of significant genes in the sub-networks and a literature mining analysis of their disease associations.

2 Methods

GenePEN identifies compact network activity alterations in functional omics datasets (here: using microarray gene expression data as an example) by casting it as a convex optimization problem. We assume a set of supervised (cases vs. controls) gene expression samples $\{\mathbf{x}_i, y_i\}_{i=1}^n$ with features $\mathbf{x}_i \in \mathbb{R}^p$ (with $p>>n$) and class labels $y_i \in \{-1, 1\}$, and an undirected graph encoding pairwise functional associations between genes (e.g., a protein-protein interaction network). We are interested in finding a *sparse* set of discriminative genes that form a large *connected* subgraph of the input graph. This is a problem of learning under *structured sparsity* (Rapaport et al., 2007; Li and Li, 2008; Bach et al., 2012; Yang et al., 2012). We adopt a penalized logistic regression approach. This involves finding weights $\mathbf{w} \in \mathbb{R}^p$ and $v \in \mathbb{R}$ that solve the program

$$\min_{\mathbf{w},v} f(\mathbf{w},v) + \lambda \Omega(\mathbf{w}), \tag{1}$$

where $f(\mathbf{w}, v)$ is the (smooth and convex) expected logistic loss

$$f(\mathbf{w},v) = \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + v))), \tag{2}$$

and $\Omega(\mathbf{w})$ is a penalty function that regularizes $\mathbf{w}$, where $\lambda \in \mathbb{R}_+$ controls the tradeoff. GenePEN implements a novel penalty function $\Omega(\mathbf{w})$ that penalizes the differences between the *absolute* values of weights of neighboring features in the graph:

$$\Omega(\mathbf{w}) = \sum_{i=1}^{p} \left[\sum_{j=1}^{p} A_{ij} |w_i| - \sum_{j=1}^{p} A_{ij} |w_j| \right]^2 + 2\Delta ||\mathbf{w}||_1^2, \tag{3}$$

where A is the (symmetric) adjacency matrix of the input graph and Δ its maximum degree, respectively, and $||\mathbf{w}||_1$ is the L_1 norm of the weight vector.

The use of *absolute* weights in the penalty function is a key difference of our method to previous related methods for biological network alteration analysis that also use a quadratic penalty function over the model weights but not their absolute values (Rapaport et al., 2007; Li and Li, 2008). The motivation for using absolute weights in the penalty function is that in linear logistic regression (as in any other linear model) the magnitude of a weight reflects the *relevance* of the corresponding feature in the solution, and the precise sign of the feature weight is irrelevant. Hence, by penalizing the absolute value of a weight, we "push" all irrelevant weights to zero, thereby keeping only the relevant attributes in the final model. This idea is implicit in the Lasso (Tibshirani, 1996), the Elastic Net (Hastie et al., 2009), and the Pairwise Elastic Net (Lorbert et al., 2010), a generalization of the Elastic Net, in which the parameter determining the trade-off between L1- and L2-regularization can be replaced to adjust the trade-off using other information (e.g., from a feature similarity matrix). The importance of using absolute weights in penalty functions for classification has been demonstrated also in other recent work (Yang et al., 2012), but in this case the proposed penalty functions were nonconvex, aggravating the efficient discovery of global optimal solutions. Our main theoretical result, which is key in achieving computational efficiency, is that the penalty function $\Omega(\mathbf{w})$ in (3) is convex in $\mathbf{w}$ (see proof and implementation details in the Supplementary Material). GenePEN solves the convex program in the TFOCS optimization framework (Becker et al., 2011), resulting in highly efficient optimization. The Matlab software implementation of GenePEN provides an easy-to-use function to train a model, taking the features x_i, the class labels y_i, the symmetric adjacency matrix A for the association graph and the regularization constant λ as input. As output, the learned weights w_i are provided, and the final feature selection amounts to choosing those w_i that are sufficiently far from zero. As a further feature of the software, example code and microarray and molecular interaction data is provided to perform a cross-validation for GenePEN with different performance statistics.

3 Results

We have evaluated GenePEN on public microarray gene expression datasets for Parkinson's disease, prostate cancer and colorectal cancer. All datasets were obtained from case-control studies with approx. balanced numbers of biological samples from patients and unaffected controls (see Supplementary Table S1). We mapped the data onto a human genome-scale protein-protein interaction (PPI) network assembled from 20 databases (see Supplementary Material for details on data collection and pre-processing). For comparison, in addition to GenePEN we applied the related methods Lasso, the Elastic Net (Hastie et al., 2009) and two variants of the Pairwise Elastic Net (Lorbert et al., 2010; Lorbert and Ramadge, 2013). All these approaches involve a single regularization parameter λ, tuned via 5-fold cross-validation estimation of the area under the ROC curve (see Supplementary Text; future work will involve computing the whole regularization path). In each cross-validation cycle, we computed the size of the largest connected component (LCC) of the induced subgraph after mapping the selected genes onto the PPI network. Since small sets of grouped features are more likely to occur by chance than larger connected components, the LCC is more favorable as a network feature grouping criterion than characteristics not accounting for size, e.g., the modularity or clustering coefficient.

Supplementary Figures S1–S3 show the relative size of the LCC as compared to the average total number of selected genes for each dataset. GenePEN extracts significantly more connected gene groups than the other methods across the whole regularization path. In a typical run, 34 out of 51 genes (67%) selected by GenePEN formed a connected subgraph in the PPI network vs. about 30% for another graph-based PEN variant (Lorbert and Ramadge, 2013) and less for the other methods. To illustrate GenePEN's biological utility using the subnetwork of grouped genes identified on the Parkinson's disease data as an example output (see Figure 1), we conducted a text-mining analysis (Glaab et al., 2012) and an over-representation analysis for differentially expressed genes (DEGs), showing that genes in the selected subgraph are enriched in significant DEGs and in positive text-mining scores with regard to the disease term "Parkinson's disease" (see Supplementary Text). Moreover, in the Supplementary Material feature selection results are evaluated on simulated data, and

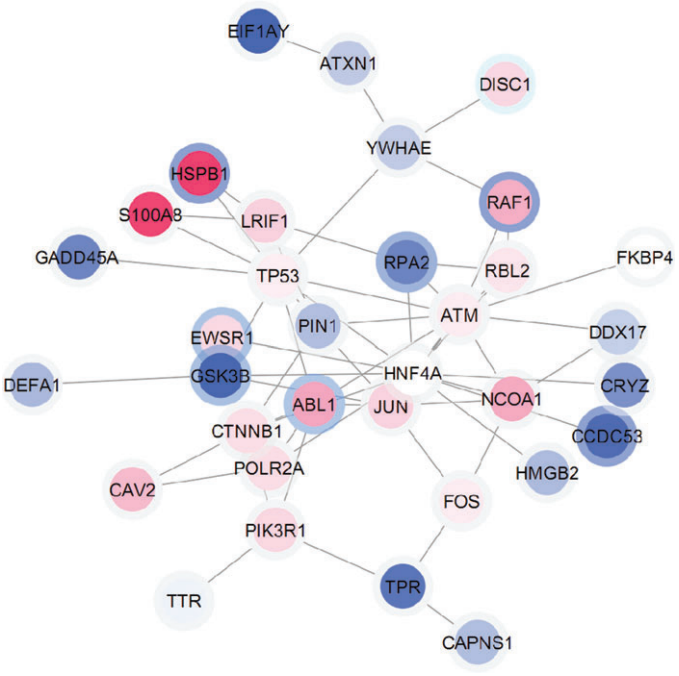

Figure 1 Visualization of the sub-network of predictive genes selected by GenePEN on the Parkinson's disease data (red=up-regulated, blue=down-regulated; darker colors represent larger alterations; color darkness is proportional to the logarithmic fold change; differentially expressed genes with FDR<0.05 are highlighted by blue node borders, see Supplementary Material for more details).

GenePEN consistently achieved similar or better performance in selecting truly differentially expressed genes in comparison to the alternative methods.

In summary, in comparison to other network-based classification methods, GenePEN identifies gene selections which share significantly more connections within a molecular interaction network while providing comparable predictive power and similar or higher detection rates for truly differential features. The Matlab sources for GenePEN, example data and code with usage instructions are freely available at http://lcsb–portal.uni.lu/bioinformatics.

References

Bach, F., R. Jenatton, J. Mairal and G. Obozinski (2012): "Structured sparsity through convex optimization," Stat. Sci., 27, 450–468.

Becker, S. R., E. J. Candès and M. C. Grant (2011): "Templates for convex cone problems with applications to sparse signal recovery," Math. Prog. Comp., 3, 165–218.

Glaab, E., J. Bacardit, J. M. Garibaldi and N. Krasnogor (2012): "Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data," PLoS One, 7, e39932.

Hastie, T., R. Tibshirani and J. Friedman (2009): The elements of statistical learning, volume 2, 2nd edition. New York: Springer.

Ideker, T. and R. Sharan (2008): "Protein networks in disease," Genome Res., 18, 644–652.

Li, C. and H. Li (2008): "Network-constrained regularization and variable selection for analysis of genomic data," Bioinformatics, 24, 1175–1182.

Lorbert, A. and P. J. Ramadge (2013): "The pairwise elastic net support vector machine for automatic fMRI feature selection," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process.

Lorbert, A., D. Eis, V. Kostina, D. M. Blei and P. J. Ramadge (2010): "Exploiting covariate similarity in sparse regression via the pairwise elastic net," in Proc. Artif. Intell. Stat., volume 9, 477–484.

Rapaport, F., A. Zinovyev, M. Dutreix, E. Barillot and J.-P. Vert (2007): "Classification of microarray data using gene networks," BMC Bioinformatics, 8, 35.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," J. R. Stat. Soc. Ser. B (Methodological), 58, 267–288.

Yang, S., L. Yuan, Y.-C. Lai, X. Shen, P. Wonka and J. Ye (2012): "Feature grouping and selection over an undirected graph," in Proc. Int. Conf. Knowl. Discov. Data Min., ACM, 922–930.

Supplemental Material: The online version of this article (DOI: 10.1515/sagmb-2014-0045) offers supplementary material, available to authorized users.