**Karan Uppal**

**Dec 15, 2010**

**1) Pattern Selection based on classification** (implemented in R):

A new feature selection algorithm based on a consensus scoring methodology was implemented. The algorithm is divided into three steps. The first two steps use the CMA package implemented in R, and the last step uses a particle swarm optimization based search strategy to find the optimal set of features that would allow maximum classification accuracy.

The user uploaded dataset is first randomly divided into distinct training and test sets based on the splitting percentage provided by the user (default: 60% train, 40%test).

In the **first step** of the feature selection process, a ranked list of features is generated by performing a t.test on the training dataset. Only the top X number of features (default 150; user defined) are retained for further analysis.

The **second** step uses one or more feature selection algorithms selected by the user to determine a consensus set of best features. The user can select t.test, f.test, kruskal.test, recursive feature elimination, random forest, Welch.test, Wilcox.test, lasso, elasticnet, and boosting. For every selected algorithm, the number of features is iteratively increased from 1 to X (default 150) in increments of delta (default 5) till there is no improvement in the 10fold cross validation accuracy for I number of iterations (default 10). This gives m subsets of top ranked features where m is the number of selected feature selection algorithms. A consensus score of a feature is defined as the number of subsets that include that feature, and its value ranges from 0 to m.



In the **final** step, the consensus subset of features selected by at least M\_min (default=100%) algorithms is then used as an input for the PSO algorithm. The PSO algorithm is designed to avoid particles from getting stuck in local optima. The search criteria (local or global search) depends on the rank of a particle. In other words, the best particle will have a narrow or local search space while the worst particle will have a much broader or a global search space controlled by the inertia variable in the velocity update equation of the PSO algorithm.

**2) Pattern Selection based on regression** (implemented in R):

A new feature selection algorithm that uses PSO based stochastic search to select optimal set of features and a SVM-regresssion method for evaluating the fitness or the mean squared error of the selected features. The user uploaded dataset is first randomly divided into distinct training and test sets based on the splitting percentage provided by the user (default: 60% train, 40%test).

The PSO algorithm is designed to avoid particles from getting stuck in local optima. The search criteria (local or global search) depends on the rank of a particle. In other words, the best particle will have a narrow or local search space while the worst particle will have a much broader or a global search space controlled by the inertia variable in the velocity update equation of the PSO algorithm.

3) Heterogeneous population:

* Split the training set into 10 sets of 1000 bootstrapped samples.
* Perform cross-validation on the nth set and use the remaining 9 sets for validation
* Get the best set of features for each set
* Get the consensus features that are selected in each set

Introduction

Bioinformatics procedures, such as hierarchical cluster analysis of samples across concentration-dependence experiments, show grouping of proteins that largely reflect these functional networks (Jones and Go, unpublished data). Consequently, a three stage process will be used to elucidate the protein functional networks and build redox proteomic models: 1) identification of differentially regulated Cys residues, 2) grouping of differentially regulated Cys residues from stage 1 into modules/clusters representing functionally related Cys residues/proteins, 3) development and visualization of redox proteomic network models.

Stage 1: We will use R/Bioconductor packages [88] to perform analyses using a variety of statistical tests, e.g., Significance Analysis of Microarrays (SAM) [89] and Linear Models for Microarray Analysis (LIMMA) [90], Recursive Feature Elimination using Support Vector Machines (RFE-SVM) [91], programs developed for microarray analysis but readily adaptable to redox proteomics data to identify differentially regulated Cys residues under different conditions. For these, all of the experimental data sets are developed with a pair-wise structure (e.g., control vs. experimental variable) so that analysis can be done both to determine effects of individual treatments and also individual treatments against all data sets. This will give both significance measures for oxidation of individual Cys and also a network structure for differentially oxidized Cys in response to each experimental variation.

Stage 2: Modules/clusters of differentially regulating Cys residues with similar expression patterns will be identified using the R/Bioconductor Weighted Gene Co-expression Network Analysis (WGCNA) package. WGCNA uses hierarchical clustering coupled with topological overlap dissimilarity measures to detect biologically meaningful modules [92]. Functional enrichment analysis of protein sets in each module will be performed using GeneTrail [93]. Previous studies on gene expression and metabolomic data have shown that the modules identified using this method are found to be functionally related [94,95]. The modules identified at this stage will be used for network analysis on per module basis in the next stage.

Stage 3: We will use both knowledge-based methods (e.g., GeneGo, Ingenuity Pathway Analysis, Cytoscape) and discovery-based methods such as WGCNA, Bayesian Networks, Differential Dependency Network (DDN) analysis, to develop and compare operational models of redox networks within each module from Stage 2 as described below.

Discovery-based methods. To complement methods based upon known pathway and genome/proteome-disease relationships, we will also apply probability and graphical-based methods to seek new network associations. This aspect is important because little work of this type has been done on redox proteomics networks, raising the possibility that unrecognized other functional redox networks exist. Methods will include WGCNA that uses a weighted correlation network analysis approach to construct networks [94, 95], , Bayesian network models based on previously developed methods to reconstruct protein regulatory networks for each module [94, 95, 96], and DDN, a probability-based method that generates local dependency models to detect topological changes in networks by comparing experimental conditions [97]. Cytoscape will be used for visualization of networks and network comparison will be performed using the CytoDDN plugin [98]. Procedures will include bioinformatics and computational approaches that we have previously applied to metabolomics data. Other bioinformatics and computational methods will be used as needed to clarify relationships, identify dominant elements within networks, and test models against multiple experimental datasets. Quantitative model development will use Bayesian network models for classification and well developed linear modeling approaches. These will be extended and tested in Aim 2 using mouse lung disease models.

Refrences

Gentleman, R.C., et al., Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004, 5(10): p. R80.

Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004;3:3

Tusher, V.G., R. Tibshirani, and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America, 2001. 98(9): p. 5116-21.

Guyon I, Weston J, Barnhill S, Vapnik V: Gene selection for cancer classification using support vector machines. Machine Learning. 2002, 46:389-422.

Langfelder, P., Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008, 9: 559.

Christina Backes, Andreas Keller, Jan Kuentzer, Benny Kneissl, Nicole Comtesse, Yasser A. Elnakady, Rolf Müller, Eckart Meese, Hans-Peter Lenhof. [**GeneTrail**—advanced gene set enrichment analysis](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933132/?tool=pmcentrez). Nucleic Acids Res. 2007, 35(Web Server issue): W186–W192.

[DiLeo MV](http://www.ncbi.nlm.nih.gov/pubmed?term=DiLeo%20MV%5BAuthor%5D&cauthor=true&cauthor_uid=22039529), [Strahan GD](http://www.ncbi.nlm.nih.gov/pubmed?term=Strahan%20GD%5BAuthor%5D&cauthor=true&cauthor_uid=22039529), [den Bakker M](http://www.ncbi.nlm.nih.gov/pubmed?term=den%20Bakker%20M%5BAuthor%5D&cauthor=true&cauthor_uid=22039529), [Hoekenga OA](http://www.ncbi.nlm.nih.gov/pubmed?term=Hoekenga%20OA%5BAuthor%5D&cauthor=true&cauthor_uid=22039529). Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. [PLoS One.](http://www.ncbi.nlm.nih.gov/pubmed/22039529) 2011, 6(10)

Mason MJ, Fan G, Plath K, Zhou Q, Horvath S. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. BMC Genomics. 2009 Jul 20;10:327.

Yu H, Zhu S, Zhou B, Xue H, Han JD., Inferring causal relationships among different histone modifications and gene expression. Genome research. 2008, 18(8): p. 1314-24.

Smith, P., Horvath, S. Eigengene networks for studying the relationships between co-expression modules. BMC Syst Biol. 2007, 1:54.