Pathway and Gene Selection with Guided Regularized Random Forests

Daniel Brumley
Advisor: Dr. Tyler Cook

Department of Mathematics and Statistics
University of Central Oklahoma

January 13, 2018

Motivation

- DNA microarray analysis allows researchers to simultaneously examine the expression levels of thousands of genes.

Motivation

- DNA microarray analysis allows researchers to simultaneously examine the expression levels of thousands of genes.
- The earliest methodological approaches focused on analyzing such data at the level of individual genes. For biological outcomes that depend on single mutations (e.g., cystic fibrosis) this is an effective approach.

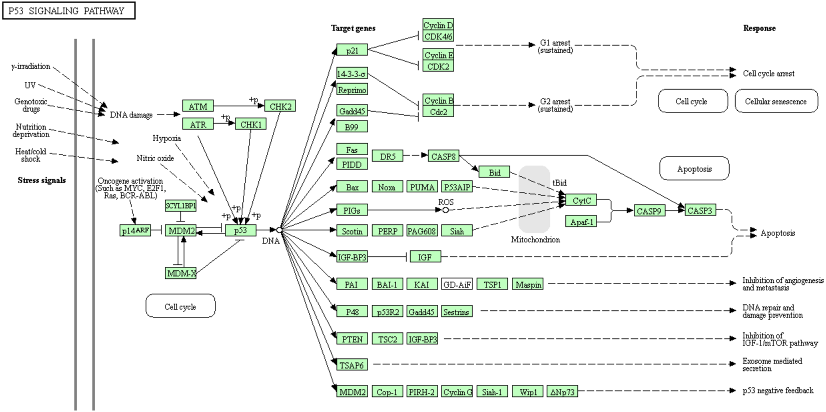
Motivation

- DNA microarray analysis allows researchers to simultaneously examine the expression levels of thousands of genes.
- The earliest methodological approaches focused on analyzing such data at the level of individual genes. For biological outcomes that depend on single mutations (e.g., cystic fibrosis) this is an effective approach.
- For diseases and disorders that are the result of complex interactions between many genes, such as cancer, this approach is less effective.

Motivation

- DNA microarray analysis allows researchers to simultaneously examine the expression levels of thousands of genes.
- The earliest methodological approaches focused on analyzing such data at the level of individual genes. For biological outcomes that depend on single mutations (e.g., cystic fibrosis) this is an effective approach.
- For diseases and disorders that are the result of complex interactions between many genes, such as cancer, this approach is less effective.
- Current cancer research improves upon the previous methodology by incorporating **genetic pathway** information into the analysis.

So, What Are Gene Pathways?



Source: KEGG Database

Random Forest Approach

Our methodology is based on that of Pang et al., who developed an approach to pathway and gene selection utilizing Random Forests:

Random Forest Approach

Our methodology is based on that of Pang et al., who developed an approach to pathway and gene selection utilizing Random Forests:

- For every pathway, construct a Random Forest model using only the genes belonging to that particular pathway.

Random Forest Approach

Our methodology is based on that of Pang et al., who developed an approach to pathway and gene selection utilizing Random Forests:

- For every pathway, construct a Random Forest model using only the genes belonging to that particular pathway.
- Rank pathways according to their out-of-bag (OOB) error rate.

Random Forest Approach

Our methodology is based on that of Pang et al., who developed an approach to pathway and gene selection utilizing Random Forests:

- For every pathway, construct a Random Forest model using only the genes belonging to that particular pathway.
- Rank pathways according to their out-of-bag (OOB) error rate.
- Within each pathway, rank the individual genes by comparing variable importance scores.

Guided Regularized Random Forests

Our methodology addresses these issues by utilizing Guided Regularized Random Forests (Deng and Runger), or GRRFs, instead.

Guided Regularized Random Forests

Our methodology addresses these issues by utilizing Guided Regularized Random Forests (Deng and Runger), or GRRFs, instead.

- In a Random Forest, splitting variables are determined by maximal information gain.

Guided Regularized Random Forests

Our methodology addresses these issues by utilizing Guided Regularized Random Forests (Deng and Runger), or GRRFs, instead.

- In a Random Forest, splitting variables are determined by maximal information gain.
- GRRF uses a regularized information gain that penalizes variables that are not already in the feature set.

Guided Regularized Random Forests

Our methodology addresses these issues by utilizing Guided Regularized Random Forests (Deng and Runger), or GRRFs, instead.

- In a Random Forest, splitting variables are determined by maximal information gain.
- GRRF uses a regularized information gain that penalizes variables that are not already in the feature set.
- The regularization parameter $\gamma \in [0, 1]$ is calculated from the variable importance scores of a preliminary Random Forest.

Breast Cancer Dataset

Following a successful round of simulation studies, the methodology was then assessed with an application to a well-known breast cancer dataset (Farmer et al.):

Breast Cancer Dataset

Following a successful round of simulation studies, the methodology was then assessed with an application to a well-known breast cancer dataset (Farmer et al.):

- The dataset consisted of three tumor classes (luminal, basal, apocrine) for 49 breast cancer patients.

Breast Cancer Dataset

Following a successful round of simulation studies, the methodology was then assessed with an application to a well-known breast cancer dataset (Farmer et al.):

- The dataset consisted of three tumor classes (luminal, basal, apocrine) for 49 breast cancer patients.
- A total of 441 pathways were utilized originating from the KEGG and BioCarta databases.

Breast Cancer Dataset

Following a successful round of simulation studies, the methodology was then assessed with an application to a well-known breast cancer dataset (Farmer et al.):

- The dataset consisted of three tumor classes (luminal, basal, apocrine) for 49 breast cancer patients.
- A total of 441 pathways were utilized originating from the KEGG and BioCarta databases.
- GRRF models were constructed for each pathway, and pathways were then ranked according to OOB error rate.

Breast Cancer Results

Pathway	Length	Gamma	Error
Glycolysis-Gluconeogenesis	68	0.10	0.022
BC Downregulated of MTA-3 in ER-negative Breast Tumors	19	0.15	0.029
BC GATA3 Participate in Activating the Th2 Cytokine Genes Expression	21	0.05	0.031
Pentose Phosphate	22	0.05	0.051
Fructose and Mannose Metabolism	39	0.05	0.055