Analyze Differently. Discover More.

DISCOVERING RELEVANT FEATURES IN HIGH DIMENSIONAL DATA

A METHOD COMPARISON CASE STUDY WITH MYXOMATOUS MITRAL VALVE DISEASE

Nathan Russell[1*], Michael Welge[1,2], Colleen Bushell[1,2], Michael Hagler[3], Nassir Thalji[3], Matthew Berry[1], Rakesh Suri[3], Loretta Auvil[1], Lisa Gatzke[1], Jordan Miller[3,4*], Bryan White[1,2*]

* Co-Corresponding Author

1 Applied Research Institute, University of Illinois at Urbana-Champaign, 2 Institute of Genomic Biology, University of Illinois at Urbana-Champaign, 3 Department of Surgery, Mayo Clinic, 4 Department of Physiology and Biomedical Engineering, Mayo Clinic

ABOUT the project

Technological advances in biological data acquisition and sequencing have enabled the identification of thousands, sometimes millions, of features per sample. Often, genotype and phenotype data are combined to identify features that are relevant to understanding a phenotype of interest.

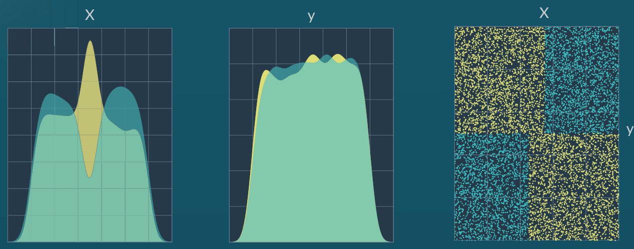
The predominant method for analyzing these data sets today typically uses univariate, or one to one comparison to identify important features in the data. However, this method is not ideal for this type of data since it is limited by assumptions of the underlying model and the data's distribution.

What We Did: Extended Random Forest (ERF)

Using high dimensional mRNA and miRNA data from Dr. Jordan Miller's research of Myxomatous Mitral Valve Disease, we tested the efficacy of the Extended Random Forest (ERF) method for computing multivariate and conditional relevance. We conclude that ERF offers different information than traditional approaches and complements traditional approaches to feature selection.

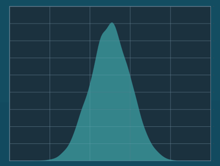
WHY is ERF better ?

ABOUT CONDITIONAL RELEVANCE



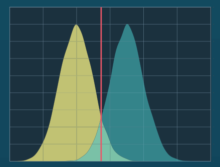
This synthetic example shows how features can have conditional relevance. The pattern depicted is a simple 'AND' operator for features X and Y. Although X and Y are clearly relevant features for prediction when combined, they aren't useful on there own. ERF is capable of detecting high order feature interactions where univariate approaches will not find X or Y.

PROPERTIES: Traditional Method



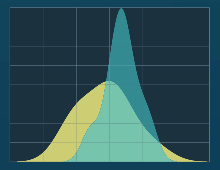
Gaussian Assumption

Two-Sample location tests (TSLTs) often have distributional assumptions or rely on asymptotic properties. If data does not match these assumptions, than the test is inappropriate.



Binary Class Limitation

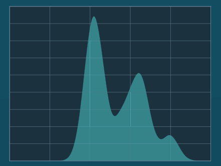
TSLTs are limited to binary classification problems. Case control studies fit this paradigm. TSLTs are limited by the assumption that the difference in means between two classes is sufficient to separate classes. This presents itself as a single discriminative partition between distributions.



Arbitrary P-Value Threshold

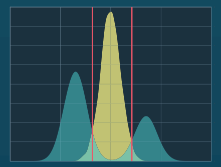
Traditional univariate feature selection procedures involve a repeated two sample test of location (TSLTs) for all candidate features and then followed by a multiple hypothesis testing correction. The subset of top features is selected by setting a hard threshold for inclusion in the set. Such thresholds are typically p-values associated with a hypothesis test.

PROPERTIES: ERF Method



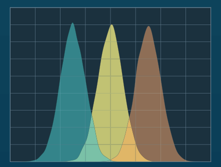
No Distribution Assumed

ERF makes no assumptions of a feature's distribution. This is sometimes preferable with real-world data where distributions are skewed or multi-modal.



Multiple Partitions

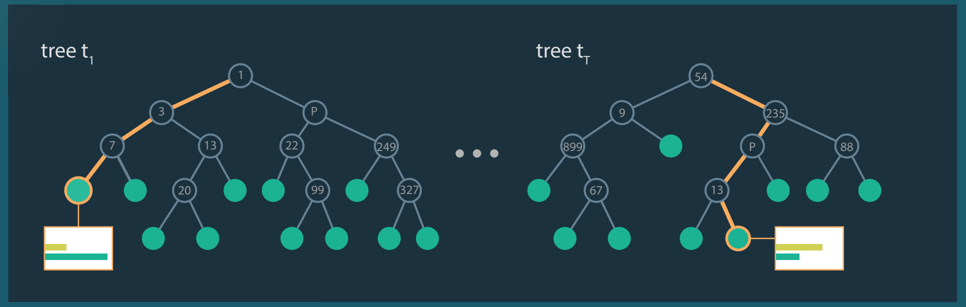
Even when there is only a single feature being considered, the recursive partitioning scheme of ERF can capture differences in distributions that are best represented as a mixture of distributions. This is sometimes the case in gene expression data sets where very high or very low expression is indicative of something different than a moderate level of expression.



Many Different Classes

The ERF framework supports multi-class classification and regression. For instance, multi-class is useful when there are multiple drug treatments in a trial or different disease subtypes in a study.

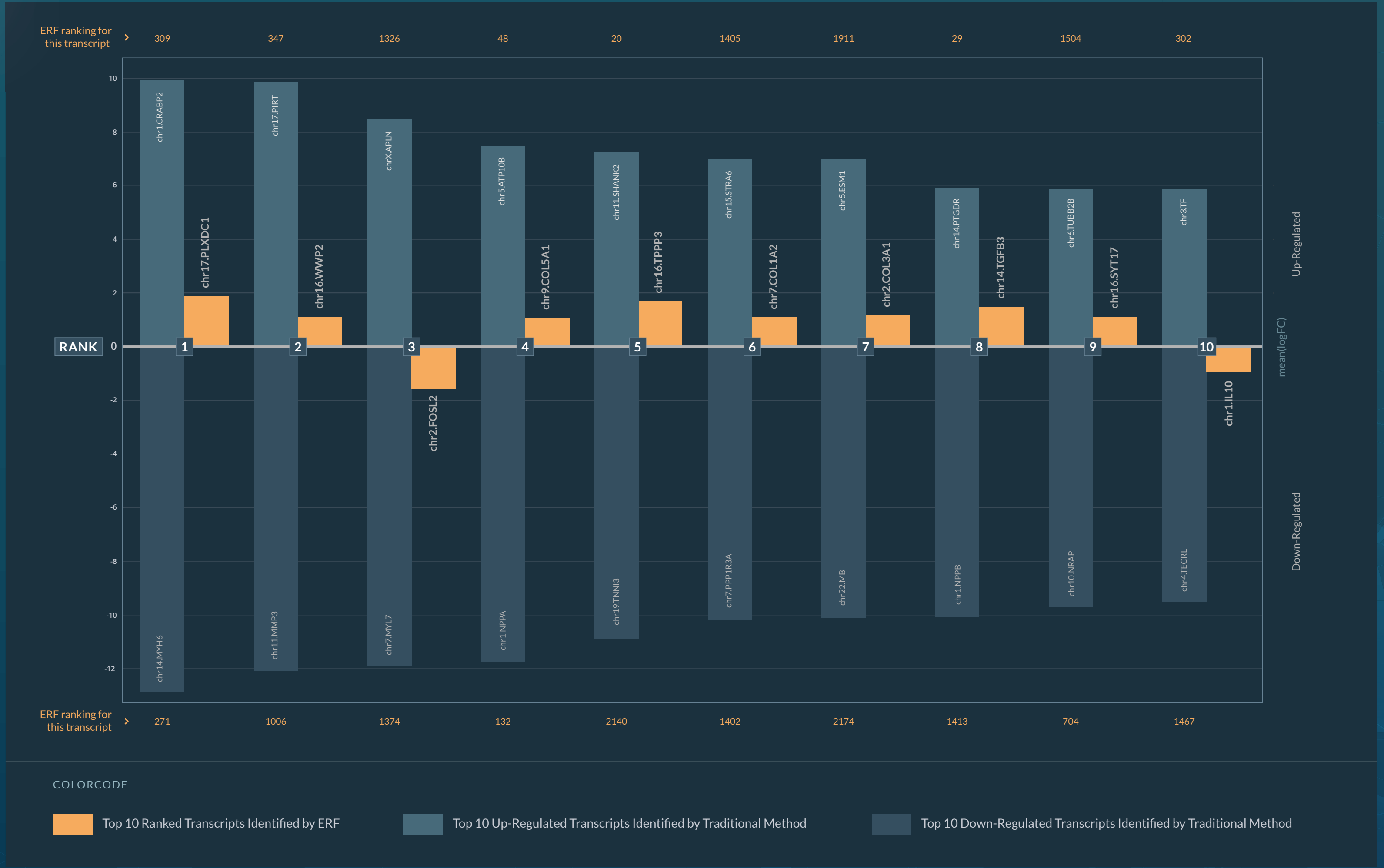
HOW ERF works



ERF generates many thousands of de-correlated decision trees that together can discover conditional relevance among sets of transcripts or other features in the data for a given classification or regression task.

ERF is a computationally intensive approach that constructs an ensemble of classification or regression trees from bootstrap samples of the data. The candidate features available at every node of the tree are a random subset of the total available features. This random sub-setting allows for trees to be fit in a lower dimension without introducing additional bias, while the averaging of these trees removes variance from the prediction. The "out-of-bag" (OOB) samples—i.e., observations left out of the bootstrap samples—are used to estimate prediction error of trees. The variable importance measure (VIM) for feature Xj, is the difference in prediction error caused by randomly permuting the Xj OOB sample values. Therefore, the VIM represents the benefit of having a feature included in a tree compared to random noise. VIMs can be ranked to identify the highly relevant features but are insufficient to declare a feature irrelevant to the disease state classification. ERF addresses this shortcoming by extending the original feature set with a permuted set of features, called shadow features. For each original feature, a shadow feature is created by randomly permuting the values from all observations. By creating these randomly permuted features X'j, we break their original associations with the response. When the permuted features are used to predict the response, the prediction accuracy decreases substantially as compared to the original feature if the original feature was informative of the response. Thus a reasonable measure for feature relevance is the likelihood that a feature has a higher mean VIM than that of the highest ranking shadow feature. This is estimated by the p-value taken from a one-sided Mann-Whitney U test between each feature and the shadow feature with the highest mean VIM.

what we LEARNED



COLORCODE

🟧 Top 10 Ranked Transcripts Identified by ERF

▨ Top 10 Up-Regulated Transcripts Identified by Traditional Method

▨ Top 10 Down-Regulated Transcripts Identified by Traditional Method

FINDINGS OF INTEREST

- None of the top 10 transcripts identified by the traditional method were in the ERF–generated list of top transcripts.
- The ERF methodology focused on the predictive contribution of a transcript and is agnostic to fold-change.
- The most relevant ERF-identified transcripts confirmed TGFβ signaling in MMVD.
- ERF identified 3 novel pathways that were intuitively relevant to MMVD.

abstract

Recent advances in biological data acquisition and sequencing technologies enable identification of thousands, sometimes millions, of features per sample [1]. A common desire is to combine datasets with unique characteristics to identify features that are relevant to understanding a phenotype of interest [2] [3]. Classical univariate testing methods are often unsuitable for such data, since they are limited by their assumptions of the underlying model and the data's distribution [4] [5].

We present the Extended Random Forest (ERF) method as one method for computing multivariate and conditional relevance in high dimensional, heterogeneous data when there is an order of magnitude more features than samples [6]. Here, we use myxomatous mitral valve disease (MMVD) as a framework to illustrate the ways in which ERF analyses can lend insights into high dimensional mRNA and miRNA data through infographics that compare statistical and computational properties.

Initially, differential gene expression in MMVD was identified using traditional univariate hypothesis testing, and features were ranked based on fold-change value and subsequent Ingenuity Pathway Analysis (IPA) of ~2,500 genes (based on cutoff criteria of fold-change > 1.5 and p < 0.05). While the ERF method identified some of the same top ranking features, numerous additional genes were more predictive of the presence of MMVD, and interestingly, none of the top ten differentially regulated genes were represented in the ten most predictive genes from the ERF analyses. Using a relevance cutoff of 95%, IPA categorization of ~450 of the most relevant ERF-identified genes confirmed the previously reported activation of TGFβ signaling in MMVD, and also identified 3 novel pathways that are intuitively relevant to MMVD. Similarly, ERF analysis of miRNA yielded novel insights due to the relative discordance between fold-change and "predictive ability" of the presence of MMVD. Collectively, our data suggest the ERF method may help to provide novel insights into multidimensional data that may lead to novel treatments and biomarkers in MMVD.

[1] Y. H. Y. B. B. Z. A. Y. Z. Pengyi Yang, "A Review of Ensemble Methods in Bioinformatics," Current Bioinformatics, vol. 5, no. 4, pp. 296-308, 2010.

[2] W. R. R. Miron B. Kursa, "Feature Selection with the Boruta Package," Journal of Statistical Software, vol. 36, no. 11, 2010.

[3] S. A. d. A. Ramon Diaz-Uriarte, "Gene Selection and Classification of Microarray Data Using Random Forest," BMC Bioinformatics, vol. 7, no. 3, 2006.

[4] B. J. N. N. M. E. v. A. D. F. E. Heidema AG, "The Challenge for Genetic Epidemiologists: How to Analyze Large Numbers of SNPs in Relation to Complex Diseases.," BMC Genetics, vol. 7, no. 23, 2006.

[5] D. J. F. K. L. K. H. B. K. T. V. E. P. Bureau A, "Identifying SNPs Predictive of Phenotype Using Random Forests," Genetic Epidemiology, vol. 28, pp. 171-182, 2005.

[6] C. B. N. R. Michael Welge, "Feature Interaction Detection with Random Forest in a High Dimensional Setting," in Individualizing Medicine Conference, Rochester, 2014.

A R I ILLINOIS APPLIED RESEARCH INSTITUTE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Funding for Visual Analytics provided by the Illinois Applied Research Institute