iterative Random Forests: stable recommendation of high-order interactions among biomolecules

Sumanta Basu [*] **, James B. Brown** [†] **, and Bin Yu** [‡]

[*]Cornell University,[†]Lawrence Berkeley National Laboratory, and [‡]University of California, Berkeley

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Integrative analysis of large, heterogeneous datasets poses a central challenge in many areas of science. Tools exist to detect important main effects and low-order interactions between pairs or small subsets of parameters; however, the detection of nonlinear, high-order interactions from real-world sample sizes has remained fundamentally unsolved. Through extensive and realistic simulation, we developed a method for detecting interactions of high-order in low-sample regimes based on Random Forests (RF) - with an order-zero increase in computational cost over the base algorithm. We regularize RF using soft dimension reduction and adaptive iterative refitting, and then decode the fitted data representation by analyzing feature usages in decision-paths. We call our approach, "iterative Random Forests", or iRF. We demonstrate the usefulness of iRF in two motivating studies: modeling enhancer sequences in Drosophila Melanogaster, and identifying chromatin-RNA interactions at alternatively spliced exons in human cells. In both settings, iRF has similar or better predictive power compared to existing approaches, and provides new insights into relationships among the features. Current challenges in the biosciences motivated the development of iRF, and the algorithm is applicable to any prediction problem in which features are well-defined and interact, approximately, in a rule-type fashion. [NEEDS UPDATE: ADD STABILITY, MORE ON BIOMOLECULES?]

random forest | dimension reduction | ensemble learning | feature selection | enhancer detection | alternative splicing

Abbreviations: RF, Random Forest; iRF, iterative Random Forest

Introduction

Biological systems arise from the actions and interactions of diverse cohorts of distinct molecular species. This is apparent in the regulation of gene expression, a core biological process necessary for all life. Gene expression is governed, in part, by the assemblage dynamics of RNA Polymerase complexes, composed of many proteins encoded by distinct genes, produced by cells in tightly regulated stoichiometric equilibria. Further, the assembly and action of an RNA Polymerase complex is regulated by interactions with cohorts of transcription factors (TFs), proteins that bind to DNA and modulate the frequency of transcription of target genes. To understand the regulation of a single gene, it is necessary to study the activities of (at least) dozens of molecules, where the action of each depends upon many others. In systems that have been richly characterized, such as the spatiotemporal expression of the gene even skipped (eve) during embryogenesis in the genetic model organism Drosophila melanogaster, TFs interact in nonlinear combinations to drive expression in precise spatiotemporal patterns (CITE). Understanding the architecture of high-order interactions is crucial for gaining insight into such biological processes, and methodologies identifying and mapping interactions impact many fields beyond developmental biology, including biomanufacturing, drug design, and precision medicine. Interactions in such systems are challenging to describe with smooth functions, as emergent properties derive from the formation of molecular machine from component molecules. While theoretical models in biophysics and biochemistry provide deep insight into the structural form of these interactions, the search for biologically important interacting elements on a genome-scale continues to be a central challenge for the biosciences.

With advances in next generation sequencing (NGS) technologies, it has become possible to survey the architecture of protein-DNA and protein-RNA interactions genome-wide, in high throughput. Databases generated by the Berkeley Drosophila Transcriptional Network Project (BDTNP) http://bdtnp.lbl.gov/ and ENCODE consortium http://encodeproject.org provide maps of TF binding events and regulatory chromatin marks for substantial fractions of the regulatory factors active in several systems early embryogenesis in the genetic model organism Drosophila melanogaster and human-derived cell lines respectively. These databases contain a previously unconceivable amount of information and hold the promise of shedding light into the complex architecture of functional regulation in bilaterian genomes. However, a central challenge lies in the fact that regulatory proteins and chromatin marks can be surveyed only one at a time, whereas they act in organized, often stereospecific groups. Hence, the discovery and mapping of interactions of unknown form and order from noisy molecular data is required to build informative models from these resources.

The BDTNP is primarily interested in understanding what molecular inputs are required to activate and regulate early body patterning genes during early development of the model organism *Drosophila Melanogaster*. In ENCODE consortium, one key area of study is splicing regulation. In splicing regulation, a large molecular machine known as the spliceosome must assemble from interactions among several RNA binding proteins (RBPs) and other large ribonucleoprotein complexes to remove introns from transcripts. Prior works in both consortia have used accurate learning algorithms [CITE: BEN ENCODE2 papers] to predict the location and activity of enhancers, splice site usage, and RNA abundance from target genes, using NGS assays of TFs, RBs, and chromatin marks as inputs. These studies, and many others in the literature, have demonstrated the incisiveness of these assays for mapping biochemical events of biological consequence. However, these models have elucidated principally main effects the marginal impacts of individual factors on complex processes already known to involve higher-order interactions (CITE). Hence, several challenges remain in pushing this frontier: *interpretable* learning machines to provide distilled insights into

Reserved for Publication Footnotes

high-order interactions that drive underlying biology. Two important bottlenecks in pursuit of this aim are: (1) the exponentially growing dimension of the search space of high-order interactions ($O(p^k)$ possible k-order interactions among p features); (2) the complex structures of machine learning algorithms, which lead to the lack of interpretability which we attribute to inherent instability in these algorithms (CITE Bin).

In this work, we take a step towards overcoming these bottlenecks and propose a novel, fast algorithm to search for important, potentially local, high-order interactions from a highly predictive learning machine. Our algorithm is built on top of Random Forest (RF), an empirically successful algorithm in genomic classification and prediction problems [CITE BIOINFORMATICS PAPERS]. Guided by the principle of stability [CITE BIN], our method, iterative Random Forest (iRF), sequentially grows feature-weighted random forests and performs soft dimension reduction of the feature space. This enables us to focus on the identification of interactions with large effects, and overcome some of the limits of small sample sizes inherent in expensive datasets (BDTNP $10M, ENCODE$ >250M). After fitting the RF, we decode the fitted data representation through importance sampling of decision rules using a recently proposed algorithm, Random Intersection Trees (RIT). This procedure enables us to extract stable high-order feature combinations prevalent in the ensemble, and the inherent structure of decision trees enables the detection of interactions that are only local an important feature for biological data, where a single molecule often performs many roles in various cellular contexts, a property known as pleiotropy (CITE). On empirical and numerical examples, we show that iRF has competitive predictive accuracy with RF, and it also extracts known and compelling novel interactions in two motivating biological problems. Beyond these applications in epigenomics and transcriptomics, we envision iRF to be useful for a large class of problems in modern biology and omics studies, e.g., Genome-Wide Association Studies (GWAS) and the detection of protein-protein interactions.

Aside from its high accuracy and broad efficacy, we chose to build on RF because the decision tree structure of its base learner matches well with underlying biological knowledge of local, combinatorial interactions (CITE Snyder). Also, for working with NGS assays where the scale of signal depends strongly on multifaceted details of the assay the invariance of RF to monotone transformation mitigates to a large extent normalization issues. We have designed our method based on extensive simulations, where the data-generating mechanisms are inspired by prior knowledge of underlying biology. Contrary to classical statistical approaches such as maximum likelihood estimates for analytically tractable models, our approach seeks to unite prior insights with the algorithm design. Simulation-driven data interrogation techniques, advocated in the past by pioneers including John Tukey, Leo Breiman and Jerrome Friedman, hold promise to enrich statisticians' arsenal for solving complex problems in data science.

Background

Since interactions among biomolecules result in emergent activities [CITE LEHNINGER], transcriptional regulation is widely believed to admit rule-like or threshold dependent interactions [CITE Biggin review]. This makes decision rule and logic based learners an attractive choice for modeling biological interactions [CITE Gerstein]. There is quantitative evidence in several model systems to suggest that a genomic region can act as an enhancer only when it is sufficiently bound by activating TFs, and sufficiently depleted for silencing complexes and chromatin marks [CITE Jasper, Anil, and Knowles]. Hierarchical decision rule-based classifiers are also well-suited to account for heterogeneity of genomic processes and the local structure imposed on feature-space by conditional active complexes and pleiotropy. These are in sharp contrast with classical statistical model structures (e.g., linear/logistic regression) in high-dimensional settings, these methods often struggle to pick up local feature interactions which are specifically of interest to biologists [CITE Biggin and Snyder].

Methods based on decision rules and their ensembles have been widely used in genome-wide association studies (GWAS) [CITE BEAM, LOGIC MODELS ETC.]. These methods, however, are designed to handle only categorical features. Measurements from NGS assays like ChIP-seq and RNA-seq are typically continuous in nature, and arbitrarily thresholding them into categories leads to the loss of predictive accuracy of the learning algorithms [CITE: ben, Biggin, Arnosti, Manollis]. More importantly, all of these methods suffer from the "curse of dimensionality", the combinatorial challenge of searching high-order interactions and the reliance of forward approaches on the informativeness of constituent lower-order interactions. In yeast, a fifth order interaction governs the activation of the alternative mating-type locus, and the marginal effects of constituent factors are of opposite sign to the action of the formed complex [CITE Jasper] a near worst-case scenario for methods reliant on the principle of marginality. The same criticism applies to decision tree based classification methods which allow for continuous features, like Rulefit, MARS, Node Harvest and forest garotte, which attempt to regress out lower order effects in the identification of high-order interactions [CITE].

Our method, iRF, leverages a recently proposed algorithm called random intersection trees (RIT) (Shah and Meinshausen, 2014) to efficiently sample from the space of all subsets of features based on their stability and predictive and predictive accuracy. For a classification problem with p binary features, RIT takes as input observations of the form $\{(Y^i, S^i)\}_{i=1}^n$, where each Y^i is a class label and each $S^i \subset \{1, \ldots, p\}$ is a set of "active" features for observation i. For instance, the class label Y^i can indicate the disease status of patient i, and S^i can be a collection of SNPS present in i. Given a class label C, the RIT algorithm repeatedly selects random subsets of observations i with $Y^i = C$ and intersects the set of active features S^i to find high-order interactions prevalent in class C. RIT produces, as an output, a collection of interactions in the form of subsets of $\{1, \ldots, p\}$ [See supplementary ** for a detailed description].

The RIT algorithm is not directly applicable to learning problems with continuous features, such as ours. However, decision rules in a tree ensemble serve as natural candidate for RIT input data, since each decision rule in a decision tree has a label and an associated set of "active" features, viz., the features used to define the rule. We exploit this structure to probe the deep decision rules in tree ensembles (Supplementary Methods). Hence, we are able to detect rule-like interactions among features without relying on the informativeness of constituent marginals. Further, RIT is very fast, and comes with theoretical guarantees for detecting stable feature interactions in the data with high probability. Using RF to localize RIT sampling, we fill a gap in the machine learning literature by decoupling the order of detectable interaction from the computational cost of discovery.

Method: iterative Random Forests (iRF)

The RIT algorithm cannot be directly applied to find interactions when the features are continuous. However, if we fit an ensemble of decision trees to the data, RIT can be used to explore the feature interactions prevalent in decision paths of the resulting decision rules. This provides a natural way to generalize RIT for classification problems with continuous features. Interestingly, applying RIT directly to all the rules in an RF leads to a very unstable estimate of interactions - slight perturbation of data leads to very different results. To encourage stability of the learned interactions, we focus only on large, pure nodes in an ensemble. To deal with the high-dimensionality of the data, we use a soft dimension reduction step, where we grow a sequence of feature weighted RF and slowly reduce dimension of the feature space.

Formally, suppose we have training data $\mathcal{D}$ in the form $\{(X^i, Y^i)\}_{i=1}^n$, containing genomic measurements of p features $X^i = (X_1^i, \ldots, X_p^i)$ (e.g., abundance of p TF proteins near an enhancer), and a binary label $Y^i \in \{0, 1\}$ indicating a genomic event of interest (e.g., whether the enhancer is active). The goal is to find subsets $S \in \{1, \ldots, p\}$ of interactions which are specifically prevalent within a class $C \in [0, 1]$ (e.g., TF proteins highly enriched or depleted only near active enhancers).

We begin by introducing feature weighted random forests. For a set of non-negative weights $w = (w_1, \ldots, w_p)$, we use RF(w) to denote a modified version of Breiman's RF. In RF(w), instead of taking a uniform random sample of features during a node split, one chooses the j^{th} feature with probability proportional to w_j. These weighted ensembles have been proposed in [CITE AMARTUNGA] under the name 'enriched random forests' and used for feature selection in genomic data analysis. With this notations, Breiman's original RF amounts to RF(w) with $w = (1/p, \ldots, 1/p)$.

The first iteration of iRF ($k = 1$) starts with Breiman's RF ($w^{(1)} := (1/p, \ldots, 1/p)$) applied on the whole data $\mathcal{D}$, and stores the importantce (mean decrease in Gini impurity) of the p features $(I_1^{(1)}, \ldots, I_p^{(1)})$. We then apply $RF(w^(1))$ separately on B bootstrap samples $\{\mathcal{D}_{(b)}\}_{b=1}^B$ of training data. From the output of each of the B RFs, we extract interactions as follows.

The output of RF(w) can be viewed as a collection of decision rules (leaf nodes in decision trees). A decision rule R is a label $\hat{Y}(R)$ assigned to a rectangular region of the feature space, and takes the form $Y = \hat{Y}(R)$ if $\{X_{i_1} > a_{i_1}, X_{i_2} < a_{i_2}, \ldots, X_{i_d} > a_{i_d}\}$, for some $i_1, \ldots, i_d \in \{1, \ldots, p\}$. Here $X_{i_1}, \ldots X_{i_d}$ are the splitting features and $a_{i(1)}, \ldots, a_{i(d)}$ are the corresponding split points used by the decision tree to form the leaf node. The combination of unique indices in $\{i_1, i_2, \ldots, i_d\} \subseteq \{1, \ldots, p\}$ is the interaction associated with a rule R, denoted as $interact(R)$. In addition, we calculate two attributes for each leaf node in an ensemble: $size(R) := \sum_{i=1}^n \mathbb{1}[X^i \in (R)]$, i.e., the number of observations falling in the leaf node R and $purity(R) := \sum_{i=1}^n \mathbb{1}[Y^i = \hat{Y}(R) \text{ and } X^i \in R] \,/\, size(R)$, the proportion of these observations whose labels match the label predicted by the tree.

Nodes with large size and high purity represent large clusters of data points in our training sample having a common class label and similar measurement of features. Hence, interactions associated with these nodes are a natural place to look for high-order interactions among features regulting the genomic event of interest. Given a tree ensemble fitted on the training samples $\mathcal{D}$, a class of interest $C \in \{0, 1\}$, and two tuning parameters $p_{leaf}, \delta \in [0, 1]$, we consider the set of nodes

$$\mathcal{R} = \left\{ R : \hat{Y}(R) = C,\ size(R) \geq p_{leaf} * n_C,\ purity(R) \geq \delta \right\} \quad [1]$$

where $n_C = \sum_{i=1}^n \mathbb{1}[Y_i = C]$ is the number of observations with label C. In order to reduce the number of tuning paramaters, we have used $\delta = 0$ for our analyses. However, in simulation studies with low signal-to-noise settings, we found that setting a larger δ can help increase the accuracy of interaction detection. With this class of large, pure nodes in hand, we apply RIT on $\{(\hat{Y}(R), interact(R)) : R \in \mathcal{R}\}$ and record the interactions obtained from RIT. For each of these interaction, we calculate the proportion of times (out of B bootstrap samples) it appears as an output of RIT, and use this proportion as a *stability score* associated with the interaction.

In the analyses of NGS datasets with a large number of features, the first iteration of iRF ($k = 1$) tend to result in interactions with low stability scores. The reason is that under uniform sampling of features during node splits, even informative features have low probability of getting selected consistently at different levels of the tree. To reduce the dimensionality of feature space without discarding features with low marginal importance, iRF iteratively grows RF(w), with feature importance (mean decrease in Gini impurity) from previous iterations used as w. This encourages RF to use important features more *stably* on decision paths, and increase the stability scores of interactions obtained by RIT. Formally, for each $k = 2, \ldots, K$, iRF grows $RF(w^{(k)})$, where $w^{(k)} = (\mathcal{I}_1^{(k-1)}, \ldots, \mathcal{I}_p^{(k-1)})$ are the feature importance from the last iterate. In our numerical and real data analyses, we observed that setting the total number of iterations K in the range $3 \sim 5$ leads to stable recovery of interactions.

The complete iRF workflow is presented in Algorithm 1.

Algorithm 1: iterative Random Forest (iRF)

Input: $\mathcal{D} = \{(X^i, Y^i)_{i=1}^n, X^i \in \mathbb{R}^p, Y^i \in \{0, 1\}\}$,$C \in \{0, 1\}\}$
Input: Tuning Parameters: (p_{leaf}, K, B)
$w^{(1)} \leftarrow (1/p, \ldots, 1/p)$
for $k \leftarrow 1$ ***to*** K **do**
  Fit $RF(w^{(k)})$ on $\mathcal{D}$
  $(I_1^{(k)}, \ldots, I_p^{(k)}) \leftarrow$ Gini Importance of $RF(w^{(k)})$
  for $b \leftarrow 1$ ***to*** B **do**
    Generate bootstrap samples $\mathcal{D}_{(b)}$ from $\mathcal{D}$
    Fit $RF(w^{(k)})$ on $\mathcal{D}_{(b)}$
    $\mathcal{R} \leftarrow$ set of large leaf nodes as defined in [1]
    $\mathcal{S}_{(b)}^{(k)} \leftarrow \text{RIT}(\{(\hat{Y}(R), interact(R)) : R \in \mathcal{R}\})$
  end
  for $S \in \cup_{b=1}^B \mathcal{S}_{(b)}^{(k)}$ **do**
    $stability(S) = (1/B) \sum_{b=1}^B \mathbb{1}\left[S \in \mathcal{S}^{(b)}\right]$
  end
  $w^{(k+1)} \leftarrow (I_1^{(k)}, \ldots, I_p^{(k)})$
end
Output: $\{S, stability(S)\}_{S \in \cup_{b=1}^B \mathcal{S}_{(b)}^{(k)}}$, for $k = 1, \ldots, K$
Output: $\{RF(w^{(k)}) \text{ on } \mathcal{D}\}$, for $k = 1, \ldots, K$

Selecting tuning parameters in iRF. iRF inherits the tuning parameters associated with its two base algorithms, viz., random forest (RF) and Random Intersection Trees (RIT). The predictive performance of RF is known to be highly resistant to the choice of tuning parameters (Breiman, 2001), so we work with the default parameter choices of RF. For the Random Intersection Tree algorithm, we work with the basic version of RIT proposed in Algorithm 1 of Shah and Meinshausen (2014), with $M = 1000$ intersection trees of depth $D = 5$.

In addition to the tuning parameters of RF and RIT, the iRF workflow introduces three additional tuning parameters - (i) p_{leaf} for controlling minimum leaf node size required for detecting candidate interactions in part **A**, (ii) number of iterations K in part **C**, and (iii) number of bootstrap samples B used in part **B**. For the examples considered in this paper, the results are fairly stable for $K = 3, 4, 5$, and $B \in (30, 100)$. The choice of p_{leaf}, however, is crucial for the detected interactions. With a larger choice of p_{leaf}, iRF picks very few interactions. With lower cutoff, however, it detects many noisy features in candidate interactions which affects the accuracy of feature selection. In the next few paragraphs, we provide some intuition about the role of p_{leaf} in the iRF workflow and discuss a stability-driven method of selecting this tuning parameter.

selecting p_{leaf} and connection to thresholded regularized regression. A principled selection procedure of p_{leaf} can be motivated by the similarity of iRF with the commonly used thresholded ridge/Lasso regressions. A fitted RF model can be viewed as an average of T additive models with indicator functions $\mathbb{1}[x \in R_{jt}]$, where these functions are adaptively chosen based on training data. Selecting p_{leaf} amounts to thresholding some terms of the additive model to zero. This problem is analogous to thresholded versions of ridge or Lasso regression, where one obtains a regularized estimate

$$\hat{\beta}(\lambda) := \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2 + \lambda\|\beta\|_q, q = 1 \text{ (Lasso)}, 2 \text{ (ridge)}$$

of the regression coefficients and the response is predicted as $\hat{Y}(\lambda) := \sum_{j=1}^{p} X_j \hat{\beta}_j(\lambda)$. Subsequently, the coordinates of $\hat{\beta}$ smaller than a threshold δ are set to zero, i.e., $\tilde{\beta}_j(\lambda, \delta) = \hat{\beta}_j(\lambda)\mathbb{1}\left[|\hat{\beta}_j(\lambda)| > \delta\right]$, for $j = 1, \ldots, p$. The final predictions are obtained as $\tilde{Y} = \sum_{j=1}^{p} X_j \tilde{\beta}_j(\lambda, \delta)$. In our problem, the RF predictions are analogous to the first stage Lasso estimates $\hat{Y} = \sum_{j}^{p} \hat{\beta}_j(\lambda)$, where the degree of regularization is governed by RF's internal tuning parameters. Selecting p_{leaf} then amounts to the selection of the thresold δ.

selecting p_{leaf} with ES-CV. In view of the above connections between iRF and thresholded regularized regressions, a possible strategy is to choose p_{leaf} using cross-validation (CV). However, even in regression problems with Lasso, CV-based model selection is known to be unstable. In this work, we adopt an estimation stability based approach (ES-CV) proposed in Lim and Yu (2015) to select p_{leaf}. In particular, we partition the training set into V blocks of equal sizes. For a given value of $p_{leaf} = p$ and each $v = 1, \ldots, V$, we leave out the samples in block v, fit iRF and threshold the terms in the additive model corresponding to the leaf nodes smaller than $p * n_y$ to zero. Then we predict the response ($\hat{Y}_v(p)$) on the whole training set based on this fitted submodel. The Estimation Stability metric $ES(p)$ proposed in Lim and Yu (2015) then takes the form

$$ES(p) = \frac{1/V \sum_{v=1}^{V} \left\| \hat{Y}_v(p) - \hat{m}(p) \right\|^2}{\hat{m}^2(p)}, \hat{m}(p) := \frac{1}{V} \sum_{v=1}^{V} \hat{Y}_v(p)$$

In binary classification, we choose p_{leaf} as the smallest local minima of $ES(p)$ which is at least as large as the cross-validated maximizer p of AUC (area under ROC curve). In Figure 1, we demonstrate the ES-CV based selection of p_{leaf} for the enhancer data analyzed in section 15. In the left, we plot ES and cross-validated AUC (over V blocks) across different values of p_{leaf}. The first local minimum is observed at $p = 0.05$. In the figure on right, we plot the test set accuracy of the submodels (only using nodes with size larger than $p_{leaf} * n_1$), for different values of p_{leaf}. We find that a choice of $p_{leaf} = 0.05$ results in only a 4% loss in predictive accuracy compared to the full model.

Grouped features and replicate assays . In many supervised learning problems with omics data, one faces the problem of drawing conclusion at an aggregated level of features at hand. The simplest example is the presence of multiple replicate assays of a single feature in the data sets, when there is neither a standard protocol to choose one assay over the other, nor is any known strategy to aggregate the assays after normalizing them individually. Similar situations arise when there are multiple genes from a single pathway in the feature sets, and one is only interested in learning interactions among the pathways and not the individual genes.

In linear regression based feature selection methods like Lasso, such grouping information among features are usually incorporated by devising suitable grouped penalties, which requires solving new optimization problems. The invariance property of RF to monotone transformation of features and the nature of intersection operation provides iRF a simple and computationally efficient workaround to this issue. In particular, one uses all the unnormalized assays in the tree growing procedure, and collapses the grouped features or replicates into a "super feature" before taking random intersections. iRF then provides interaction information among these super features, which could be used to achieve further dimension reduction of the interaction search space.

Case Study I: Learning Architecture of enhancer elements in Drosophila

In animals, definitive epigenetic signatures of enhancer elements have been challenging to identify - the best prediction tools offer weak positive predictive power at genome-scales and are not sufficiently accurate to conduct in silico enhancer annotation based on ChIP-seq data. In the early Drosophila embryo, a small cohort of 40 transcription factors drive body patterning (for a review, see Rivera-Pomar and Jckle (1996)). Hence, fly development offers a simplified model system in which to study the relationship between transcription factor binding and tissue-specific enhancer activity.

We studied the DNA binding patterns of 22 of these factors, as well as chromatin marks during embryonic stages 4 and 5. Specifically, we collected previously tested elements using in situ enhancer assays. These assays test genomic elements in ectopic reporter assays stably integrated into a dedicated location in the genome. A sequence is scored as positive in this assay (an active enhancer) if it drives patterned expression at one of more developmental stages. We selected enhancers active during the period of the blastoderm (stages 4-6, 80 minutes) - one of the best studied periods for gene regulation in any organism. In total, 7705 genomic elements were studied by Kvon et al. (2014) and 110 by Fisher et al. (2012). We added to this 172 additional (new and unpublished) elements from the Celniker lab of Lawrence Berkeley National Laboratory available from the Berkeley Drosophila Genome Project (BDGP at http://fruitfly.lbl.gov). Embryos were collected between 0 and 16 hours old (approximately uniform random age distribution from developmental stages 0 through 16), and between one hundred and one thousand animals were imaged to assess expression status for each construct. Among these el-

ements, 731 were found to drive expression in the embryo at stages 4-6.

To obtain insight into why these elements drive patterned expression, we studied 35 ChIP assays on 23 transcription factors (expressed in patterns, data from Berkeley Drosophila Transcription Network Project, http://bdtnp.lbl.gov, MacArthur et al. (2009); Li et al. (2011)) and also data for 45 histone modifications Li et al. (2014). For each of these data tracks, we computed maximal score over each genomic element, and these became input features. Additionally, we computed the average and maximal average score along sliding windows of lengths 200, 500 and 1000 nucleotides. The windowing approach is useful because element lengths vary from 400 nt to ¿3000 nt, and it is likely that for many elements only a small sub-sequence drives patterned expression. We used these measurements to form a set of potential features for the enhancer activity status (active or not, binary) of each genomic element (total dimensionality of our feature space XXX). We randomly divided the dataset into balanced training and test sets, and applied iRF with B = 30, K = 3 and a leaf node cutoff of 35. The tuning parameters in RF were set to default and 1000 binary random intersection trees of depth 5 were grown to capture candidate interactions. We report the interactions with stability scores of at least 40% (equivalent to a p-value cutoff of approximately XXXX). The results of prediction and feature selection of iRF are displayed in Figure 5.

In this prediction problem, we achieve a stable AUROC of 0.82 throughout the first three iRF iterations. RIT applied on RF (iteration 1) did not pick up any feature interactions with stability score more than 40

Case Study II: Epigenomic Landscape of Alternative Splicing in Human

In eukaryotes, alternative splicing of primary mRNA transcripts is a highly regulated process by which multiple distinct mRNAs are produced by the same gene. There are believed to be as many as 1,500 RBPs in the human genome (Gerstberger et al., 2014) - a remarkable number accounting for nearly 8

The ENCODE consortium has collected extensive genome-wide data on both chromatin state and gene expression in the human-derived erythroleukemia cell line K562 (Consortium et al. (2012), http://encodeproject.org). To identify important interactions at the basis of chromatin mediated splicing, we used splicing rates (Percent-spliced-in, PSI values, Pervouchine et al. (2016)) from ENCODE RNA-seq data, along with ChIPseq and ChIP-chip for chromatin marks and transcription factor binding events (253 ChIP assays on 107 unique transcription factors and 11 histone modifications, https://www.encodeproject.org/). For each ChIP assay, we computed the maximal score over the genomic region corresponding to each exon. This yielded a set of p = 270 features for our analysis. We took our response to be a thresholded function of the PSI values for each exon. Only internal exons with high read count (at least 100 rpkm) were used in downstream analysis. Exons with Percent-spliced-in index (PSI) above 70

We find that iRF achieved slightly better predictive accuracy compared to RF (Figure 6A), as measured by the area under ROC curve (AUC). iRF identified highly stable interactions between H3K36me3, as expected, a number of novel interactions involving other chromatin marks, as well as post-translationally modified states of RNA Pol II (Figure 6B). In particular, serine 2 phosphorylation of Pol II is highly connected to associated marks. Examining surface maps for important pairwise interactions (Figure 6C) reveals effect-sizes of 20-40

Numerical Experiments

In this section we demonstrate the advantage of iRF over RF on simulated datasets.

Experiment I: Dimension Reduction with Iterative Learning. Motivated by the stereospecific nature of interactions among biomolecules, we generate datasets from Boolean rule-based models. In every simulation setting, $p = 50$ features $(X_1, \ldots, X_p)$ follow independent standard Cauchy distribution.

The binary response variable Y is generated in the three settings (AND, OR and XOR) as follows:

$$\begin{aligned}
\textbf{OR:} \quad & Y = \mathbb{1}\left[X_1 > t_1 \,|\, X_2 > t_1 \,|\, X_5 > t_1 \,|\, X_8 > t_1\right] \\
\textbf{AND:} \quad & Y = \mathbb{1}\left[(X_1 > t_2 \,\&\, X_2 > t_2) \,|\, (X_5 > t_2 \,\&\, X_8 > t_2)\right] \\
\textbf{XOR:} \quad & Y = \mathbb{1}\left[(X_1 > t_3 \,\&\, X_2 < -t_3) \,|\, (X_1 < -t_3 \,\&\, X_2 > t_3)\right. \\
& \left. \;\; |\, (X_5 > t_3 \,\&\, X_8 < -t_3) \,|\, (X_5 < -t_3 \,\&\, X_8 > t_3)\right]
\end{aligned}$$

Note that the model family "OR" does not explicitly code for interactions among biomolecules, rather represents alternative mechanisms driven by univariate features. From a modeling perspective, this gives rise to a non-additive main effect.

To ensure that the two classes $Y = 1$ and $Y = 0$ are not extremely unbalanced, we set $t_1 = 3.2$, $t_2 = 0.2$ and $t_3 = 0.4$. Training Samples of size $n \in \{25, 50, 75, \ldots, 500\}$ are simulated from the above generative models and the performance of different iterations of iRF (iteration $1 \equiv$ RF) in feature selection and prediction (on held-out test sets of equal size) are recorded and reported over 20 replicates. To assess the inherent complexity of the underlying problem, predictive accuracy of an oracle classifier, a Random Forest using only the features $\{1, 2, 5, 8\}$ was also investigated on every simulated dataset. The results are summarized in Figure 2.

The tuning parameters of Random Forests were set at default and $M = 1000$ binary random intersection trees of depth 5 were grown in each case. We used $B = 30$ bootstrap replicates and the candidate interactions used for random intersection trees were taken only from nodes with at least 10% of the training observations.

The results on the second column of Figure 2 show that both RF and iRF have similar AUROC when the sample size exceeds a certain threshold, which changes depending on the structures of the rules. XOR rules seem harder to learn than AND and OR rules with comparable sample capacity, which is expected since none of the features in XOR has strong marginal effect on Y.

The third and the fourth columns of Figure 2 demonstrate the proportion of true recovery and the total number of interactions detected (on average) for each of the three model classes. As in the case of predictive accuracy, the figures clearly show that for smaller sample size, different iterations of iRF perform better than RF. However, with larger sample size, RF also performs comparatively well as iRF.

Experiment II: Iteration as regularization - Bias-Variance Analysis. The soft dimension reduction technique can viewed as a form of regularization on the base RF learner, since it restricts the form of functions RF is allowed to fit in a probabilistic manner. To gain insight into the working of iRF as a regularizer, we conducted a bias-variance analysis of iRF predictions. Since the bias and variance components in a classification problem with $0-1$ loss function are not independent, we conducted the analysis on a regression problem.

In this problem, we generated samples from a linear regression model $Y = \sum_{j=1}^{p} X_j\beta_j + \epsilon_j$, with $p = 100$, $\epsilon \sim N(0,1)$, $\beta_1 = \ldots = \beta_10 = 0.5$ and $\beta_j = 0$ for $j > 10$. We generated $n = 200$ training samples from the above model and reported average bias, average variance, mean-squared error (MSE) and relative mean-squared-error (RMSE) of RF and iRF predictions in a held-out test set of equal size. The results, averaged over 500 replications, were reported in Figure 3. In each replicate, we simulated the errors (ϵ_j) for the training set. The average bias is estimated as the difference between the mean prediction over 500 replicates and the expected response on the test set. The average variance is estimated as the variance of the 500 predictions for each of the $n = 200$ training samples.

The first two plots show that the average bias of the learner decreases with iteration, while the average variance increases. The third plot shows that the overall MSE decreases with iteration, although the decrease is minimal after the first two iteration. The bias reduction effect of iRF with ensemble of weighted decision trees is similar to the bias reduction effect of adaptive lasso, as mentioned in (Meinshausen, 2009; Zou, 2006).

Experiment III: Computational cost of detecting high-order interaction We use the two datasets from our case studies to demonstrate the computational advantage of iRF for detecting high-order interactions from high-dimensional data. Rulefit3 serves as a benchmark, which has competitive prediction accuracy to RF and also comes with a flexible framework for detecting nonlinear interactions hierarchically, using the so-called "H-statistic" (Friedman and Popescu, 2008). We show that for moderate to large dimensional datasets typically encountered in omics studies, the computational complexity of calculating H-statistic increases rapidly, while the computation time of iRF grows far more slowly with dimension.

We fit iRF and rulefit on balanced training samples from the two datasets, enhancer identification (7705 samples, 86 features) and alternative splicing ($24k$ samples, 301 features) using subsets of p randomly selected features, where $p \in \{10, 20, \ldots, 80\}$ for the enhancer data and $p \in \{50, 100, \ldots, 300\}$ for the splicing data. We ran rulefit with default parameters, generating null interaction models with 10 bootstrap samples and looked for higher order interactions among features whose H-statistics are at least one null standard deviation above their null average. The current implementation of rulefit only allows H-statistic calculation for interactions of up to order 3, so we do not assess higher order interactions. We run iRF with $B = 10$ bootstrap samples, and the same tuning parameter specifications as described in the case studies section. We plot the run time (in minutes) and the area under ROC curve (AUROC) for different values of p in Figure 4.

The two plots on the left panel show that the runtime for rulefit's interaction detection increases rapidly as p increases, while the increase is almost linear for iRF. For the splicing dataset, rulefit ran for over 24 hours for $p = 250$ while it took only a little above an hour to run iRF on $p = 300$. We note that the current implementation of Rulefit3 uses an optimized executable program, while iRF uses a far less efficient R implementation. On the other hand, iRF is run in parallel on 10 servers, while the current implementation of rulefit does not allow parallelization. The search space of rulefit is restricted to all possible interactions of order 3, while iRF searches for arbitrarily high-order interactions, leveraging the deep structure decision trees in RF. The linear vs. polynomial growth of computing time is not an optimization issue, it is merely an artefact of the exponentially growing search space of high-order interactions. [check edits].

Discussion

Systems governed by nonlinear interactions are ubiquitous in biology. Here, we focused on transcriptional and co- and post-transcriptional regulation. We identified known and novel interactions in early zygotic enhancer activation in the Drosophila embryo, and posit new high-order interactions in splicing regulation in a human-derived system. These interactions are stably detected, contribute to predictive power in held out test data statistically they are valid.

The direct validation and assessment of complex interactions in biological systems will be challenging, but new tools are becoming available for targeted genome engineering that promise to enable interrogation through wet-lab studies. For instance, the CRISPR system has been modified for the targeted manipulation of post-translational modifications to histones (Hilton et al., 2015), hence, it may soon to possible test if modifications to distinct residues at multivalent nucleosomes, e.g. H3K20me1 and H3K9ac, function in a non-additive fashion in splicing regulation.

Population genetics also offers exciting opportunities for the application of the iterative learning approach we have implemented in iRF. Genome Wide Association Studies (GWAS) seek to identify genetic variants that contribute to individual traits. With the exception of Mendelian disease, risk alleles function in complex networks that are often challenging to dissect with extant statistical procedures, particularly when the effects of multiple variants are non-additive, or epistatic (for a recent review see Szymczak et al. (2009)). The extent to which non-additive effects drive human phenotypes is still debated [CITE], but family studies have made it clear that risks for many complex human diseases derive from non-additive interactions between multiple genes (Brown et al. (2014) and for a review see Manolio et al. (2009)) including type I diabetes (Clayton, 2009), coronary artery disease (Schunkert et al., 2011), and many others (for a review see McCarthy et al. (2008)). Non-additive effects may be pervasive in human disease if most effects are high-order, we would not know it given the limitations of current computational and statistical methods combined with available GWAS sample sizes. Studies in tractable model organisms have demonstrated that strong epistatic dependencies are the rule, rather than the exception in higher metazoans (Mackay, 2014; Aylor et al., 2011). Geneticists are particularly accustomed to challenges associated with epistasis and non-additive interactions indeed this is exploited in synthetic lethal screens: some variants with no detectable phenotype in a wild-type background are lethal when in combination with another allele. In such settings, forward approaches are faced with the challenge that marginal effects are only observable in the susceptible subset of the population, which can be quite small and difficult or impossible to identify a priori. Methods that require the exhaustive enumeration of low-order terms (the principle of marginality), such as H-statistics in Rulefit3, are computationally infeasible when the number of candidate variants is large which is virtually always the case. We propose that a grand challenge for human genetics is the construction of statistical and computational procedures that can identify epistatic networks at, or near, the same efficiency as single risk alleles from population studies. Application of iRF to these heterogeneous, low effect-size, N ¡¡ P regimes holds promise as an area for future work, and progress may enable the emergence of Genome Wide Epistasis Studies (GWES) where the target of analysis is the discovery of epistatic networks rather than individual alleles. Validation in this setting may also be more straightforward, as accurate inferences should enhance our understanding of the heredity

for complex diseases, as measured by the fraction of explained heritability.

However, numerous challenges await: iRF currently handles data heterogeneity only implicitly. It will be useful to add tracking of individual observations. This amounts to simultaneously boosting observations and features at each node. Such a strategy would further localize our feature selection procedure. Additionally, we currently generate surface maps naively from the data we need to make better use of the information stored in the underlying ensemble to decode the density estimate implicit in the smoothed feature splits.

To date, machine learning has been driven largely by the need for accurate prediction. In science, however, we aim to understand why prediction is possible to understand the mechanics that underlie natural and artificial systems. Introspective learning procedures may ultimately be widely adaptable to diverse algorithmic and computational architectures, and may broadly improve the interpretability and informativeness of learning machines.

ACKNOWLEDGMENTS. We thank Peter Bickel for helpful discussion and comments, the laboratory of Sue Celniker for conducting experiments on enhancer elements, Taly Arbel for preparing Drosophila datasets, the laboratory of Roderic Guigo for conducting RNA-seq experiments and Dmitri Pourvechine for quantification of exon splicing rates. JBB and BY were supported by Grants ** and SB was supported by Grant **.

References

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

J. H. Friedman and B. E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, pages 916–954, 2008.

C. Lim and B. Yu. Estimation stability with cross validation (escv). *Journal of Computational and Graphical Statistics*, (just-accepted), 2015.

N. Meinshausen. Forest garrote. *Electronic Journal of Statistics*, 3:1288–1304, 2009.

R. D. Shah and N. Meinshausen. Random intersection trees. *The Journal of Machine Learning Research*, 15(1):629–654, 2014.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Figures and Tables

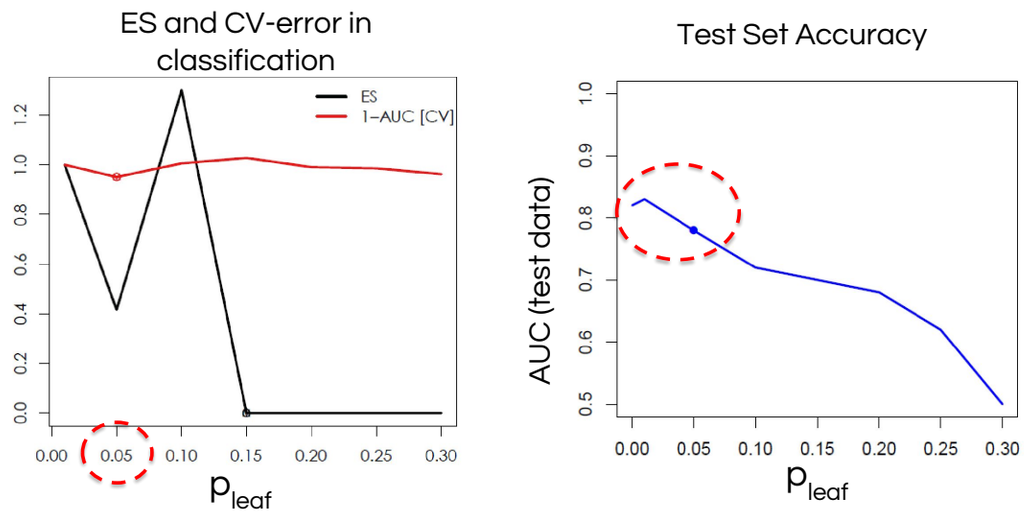

Fig. 1: Tuning parameter Selection using ES-CV on the enhancer dataset described in Section 15. In any single iteration of iRF, the tuning parameter p_{leaf} can be chosen as the first minimizer on the ES path which is larger than or equal to the maximizer of AUC.

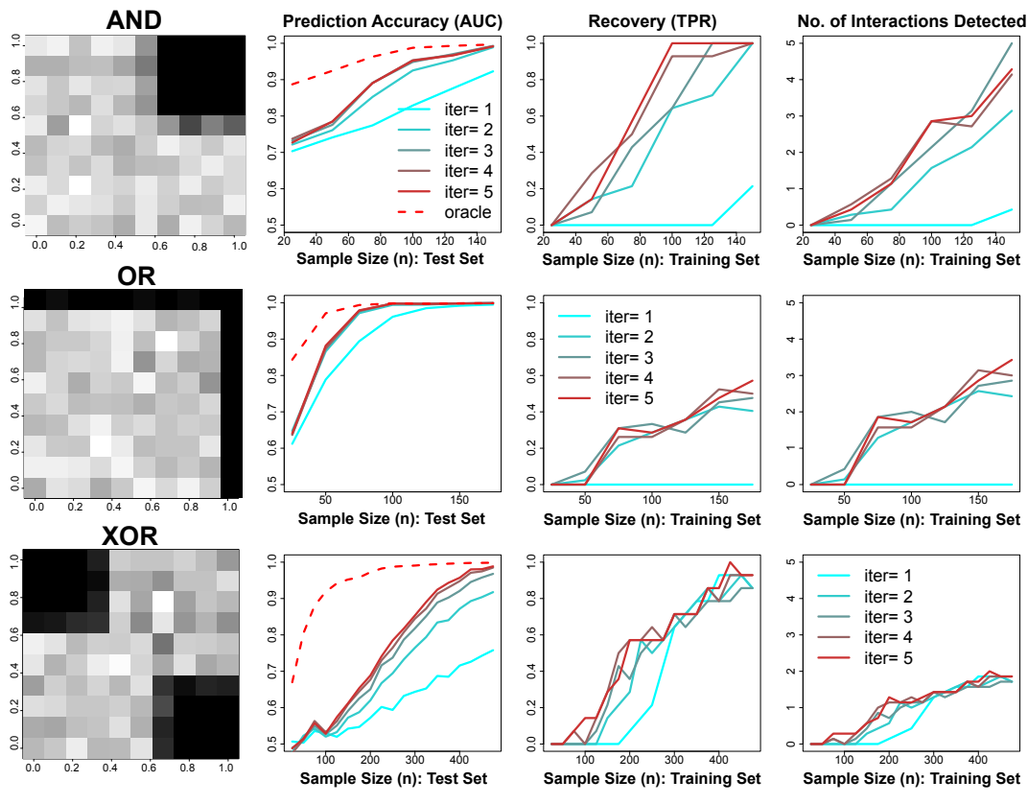

Fig. 2: Accuracy of Prediction, feature selection and number of discovered interactions in data sets simulated with rule-based interactions

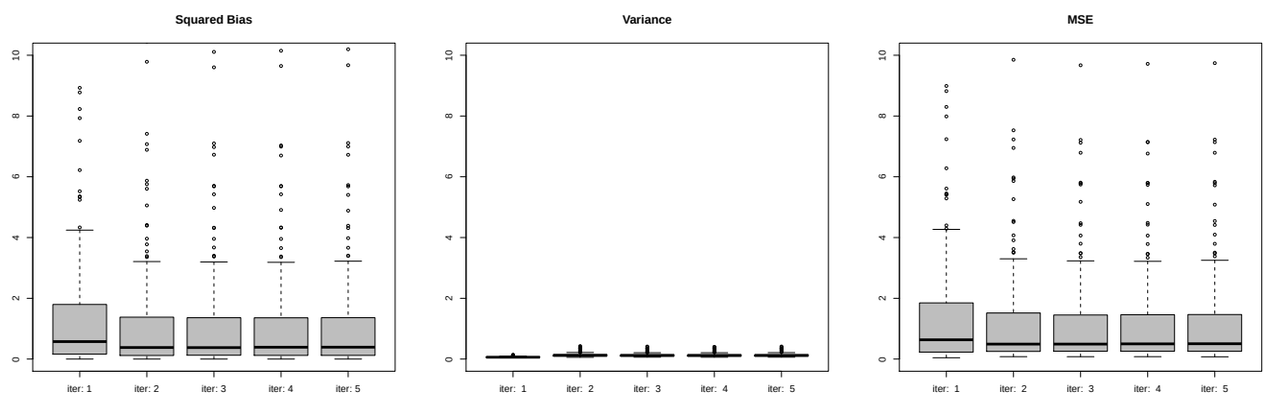

Fig. 3: Bias, Variance and MSE of different iterations of iRF in a linear regression model. The results indicate that iRF tends to decrease the bias of RF and increase the variance, but gains in overall mean squared error (MSE).

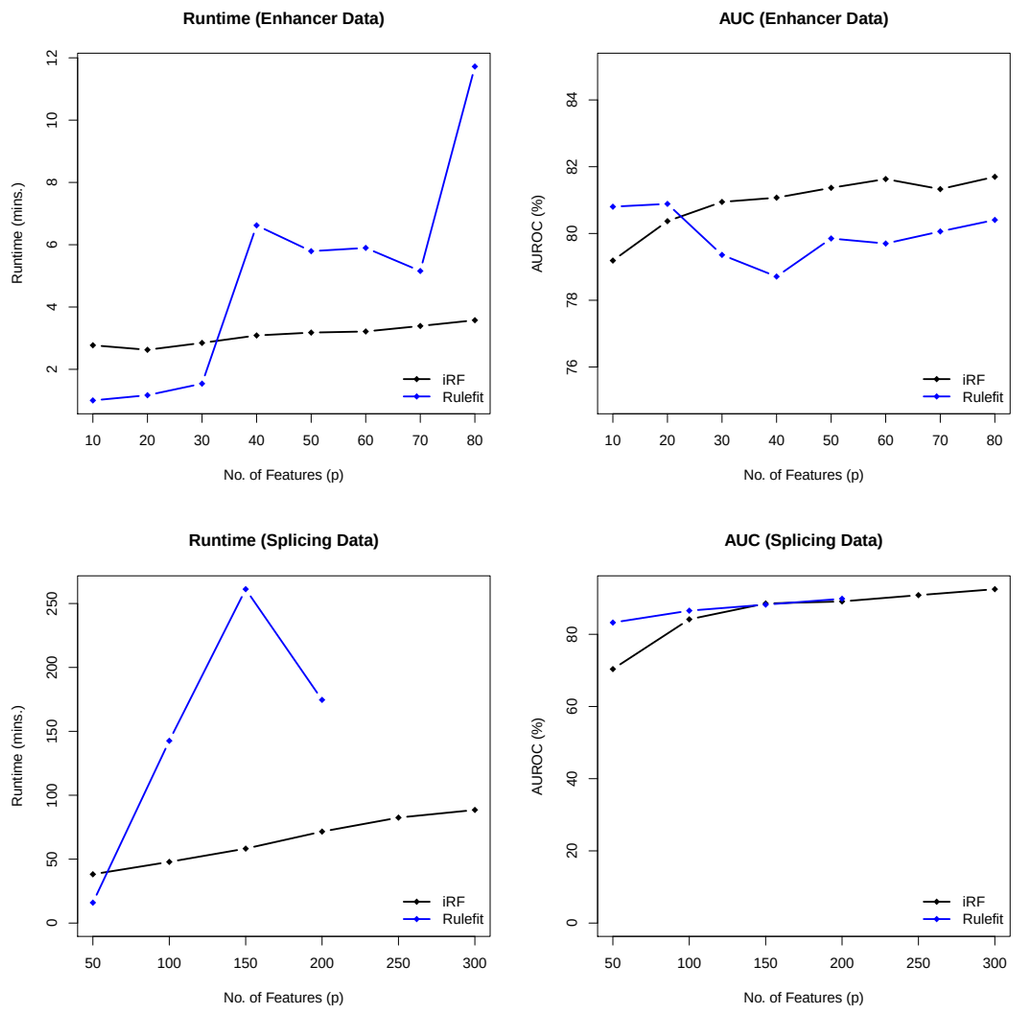

Fig. 4: Runtime (in minutes) and test set AUROC of RF and rulefit on the enhancer data and the splicing data, as the number of features in the model increases. As shown from the plots on the left, runtime of rulefit increases much faster with number of features in the model, while the increase for iRF is almost linear in p (For splicing data, rulefit is still running after 24 hours for $p = 250$) The plot on the right shows that the two methods provide comparable predictive accuracy for different number of features in the model.

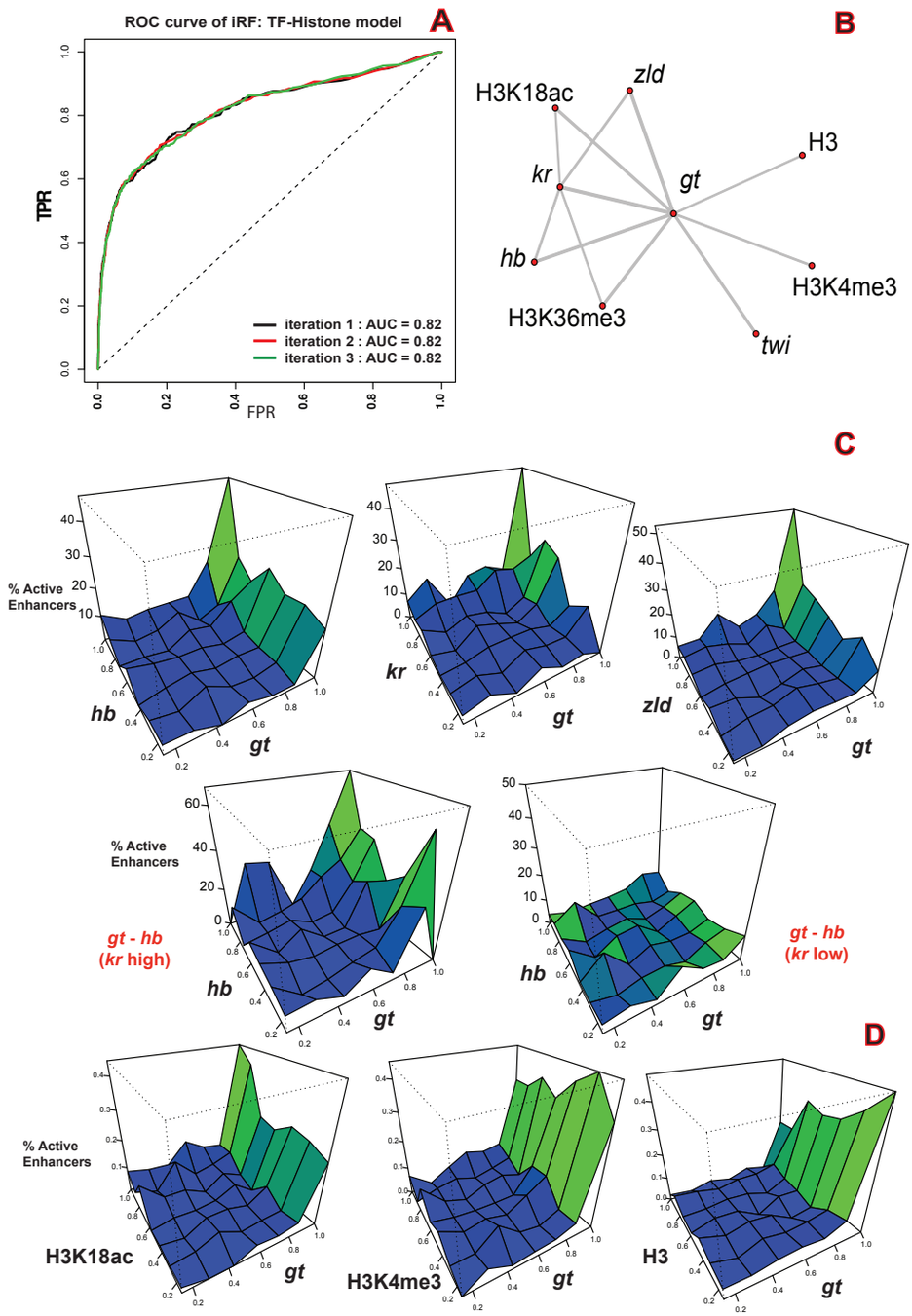

Fig. 5: [**A**]: Accuracy of iRF in predicting active enhancers with TF and histone modification data. iRF maintains same AUROC as RF. [**B**]: interactions detected by iRF after 3 iterations. Some known interactions among transcription factors *gt*, *kr* and *hb* detected. iRF also captures recently discovered interacting roles of master regulator *zld*. In addition, some novel interactions among TF and histone modifications detected. [**C**]: Surface maps demonstrating proportion of active enhancers across different pairs of TFs (used only held-out test data). Structures of response surfaces indicate presence of AND rule-type interactions. [**D**]: Surface maps demonstrating TF-histone interactions.

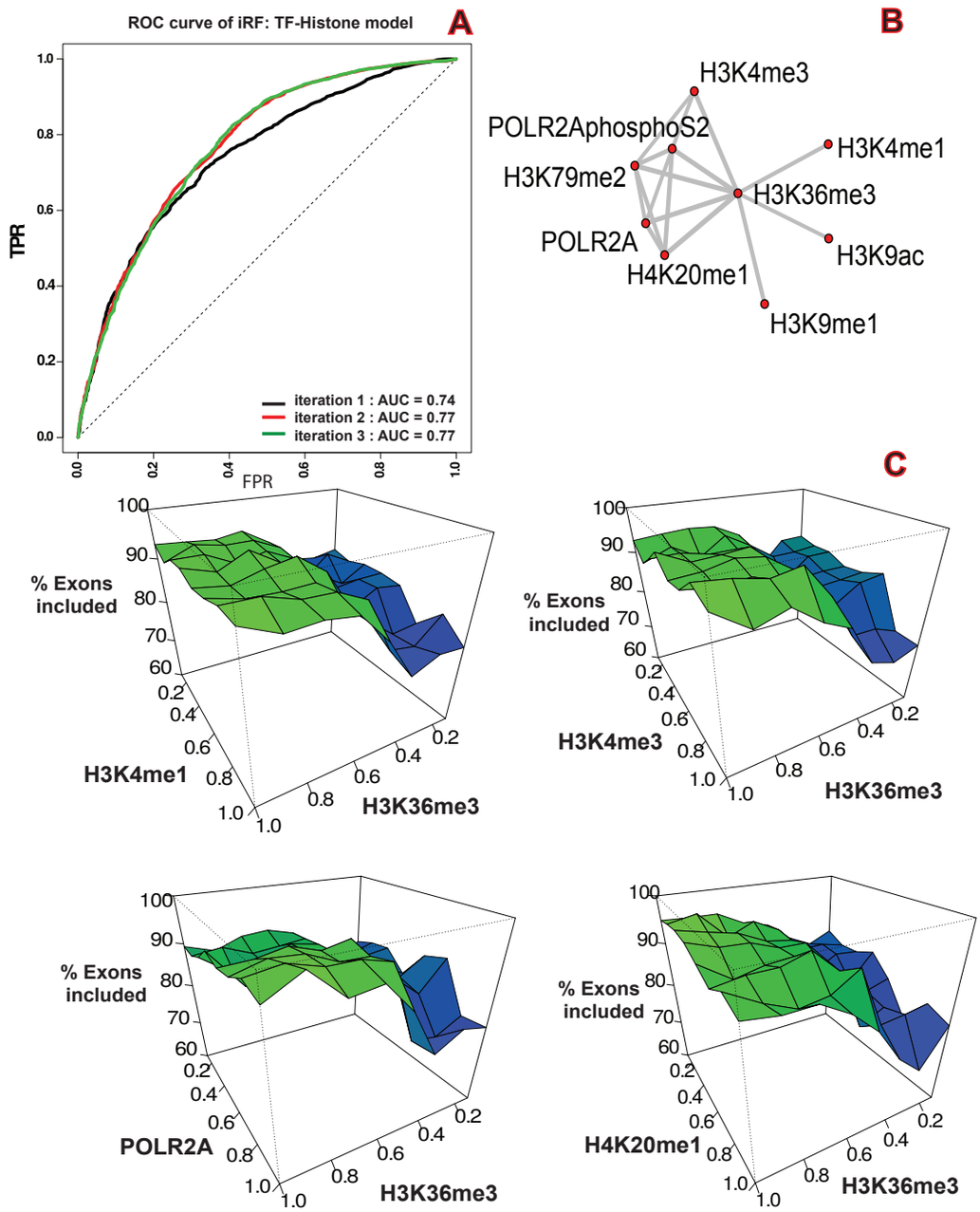

Fig. 6: [**A**]: Accuracy of iRF in classifying included exons from excluded exons in held-out test data. iRF shows 3% increase in AUROC over RF. [**B**]: interactions among TF and histones detected by iRF. [**C**]: Surface maps of excluded exons plotted across different histone modifications. The structure of response surfaces indicate presence of rule-like structures.