Iterative random forests to discover predictive and stable high-order interactions

Sumanta Basu[a,b,c,1], Karl Kumbier[d,1], James B. Brown[c,d,e,f,2], and Bin Yu[c,d,g,2]

[a]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853; [b]Department of Statistical Science, Cornell University, Ithaca, NY 14853; [c]Data Driven Decisions Department, Preminon LLC, Antioch, CA 94531; [d]Statistics Department, University of California, Berkeley, CA 94720; [e]Centre for Computational Biology, School of Biosciences, University of Birmingham, Edgbaston B15 2TT, United Kingdom; [f]Molecular Ecosystems Biology Department, Biosciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; and [g]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720

Contributed by Bin Yu, December 21, 2017 (sent for review June 23, 2017; reviewed by Michael M. Hoffman and Daniel Jacobson)

Genomics has revolutionized biology, enabling the interrogation of whole transcriptomes, genome-wide binding sites for proteins, and many other molecular processes. However, individual genomic assays measure elements that interact in vivo as components of larger molecular machines. Understanding how these high-order interactions drive gene expression presents a substantial statistical challenge. Building on random forests (RFs) and random intersection trees (RITs) and through extensive, biologically inspired simulations, we developed the iterative random forest algorithm (iRF). iRF trains a feature-weighted ensemble of decision trees to detect stable, high-order interactions with the same order of computational cost as the RF. We demonstrate the utility of iRF for high-order interaction discovery in two prediction problems: enhancer activity in the early *Drosophila* embryo and alternative splicing of primary transcripts in human-derived cell lines. In *Drosophila*, among the 20 pairwise transcription factor interactions iRF identifies as stable (returned in more than half of bootstrap replicates), 80% have been previously reported as physical interactions. Moreover, third-order interactions, e.g., between *Zelda* (*Zld*), *Giant* (*Gt*), and *Twist* (*Twi*), suggest high-order relationships that are candidates for follow-up experiments. In human-derived cells, iRF rediscovered a central role of H3K36me3 in chromatin-mediated splicing regulation and identified interesting fifth- and sixth-order interactions, indicative of multivalent nucleosomes with specific roles in splicing regulation. By decoupling the order of interactions from the computational cost of identification, iRF opens additional avenues of inquiry into the molecular mechanisms underlying genome biology.

high-order interaction | random forests | stability | interpretable machine learning | genomics

High-throughput, genome-wide measurements of protein–DNA and protein–RNA interactions are driving new insights into the principles of functional regulation. For instance, databases generated by the Berkeley *Drosophila* Transcriptional Network Project (BDTNP) and the ENCODE consortium provide maps of transcription factor (TF) binding events and chromatin marks for substantial fractions of the regulatory factors active in the model organism *Drosophila melanogaster* and human-derived cell lines, respectively (1–6). A central challenge with these data lies in the fact that chromatin immunoprecipitation sequencing (ChIP-seq), the principal tool used to measure DNA–protein interactions, assays a single protein target at a time. In well-studied systems, regulatory factors such as TFs act in concert with other chromatin-associated and RNA-associated proteins, often through stereospecific interactions (5, 7); for a review see ref. 8. While several methods have been developed to identify interactions in large genomics datasets, for example refs. 9–11, these approaches either focus on pairwise relationships or require explicit enumeration of higher-order interactions, which becomes computationally infeasible for even moderate-sized datasets. In this paper, we present a computationally efficient tool for directly identifying high-order interactions in a supervised learning framework. We note that the interactions we identify do not necessarily correspond to biomolecular complexes or physical interactions. However, among the pairwise *Drosophila* TF interactions identified as stable, 80% have been previously reported (*SI Appendix*, section S4). The empirical success of our approach, combined with its computational efficiency, stability, and interpretability, make it uniquely positioned to guide inquiry into the high-order mechanisms underlying functional regulation.

Popular statistical and machine-learning methods for detecting interactions among features include decision trees and their ensembles: CART (12), random forests (RFs) (13), Node Harvest (14), Forest Garrote (15), and Rulefit3 (16), as well as methods more specific to gene–gene interactions with categorical features, such as logic regression (17), multifactor dimensionality reduction (18), and Bayesian epistasis mapping (19). With the exception of RFs, the above tree-based procedures grow shallow trees to prevent overfitting, excluding the possibility of detecting high-order interactions without affecting predictive accuracy. RFs are an attractive alternative, leveraging high-order interactions to obtain state-of-the-art prediction accuracy. However, interpreting interactions in the resulting tree ensemble remains a challenge.

We take a step toward overcoming these issues by proposing a fast algorithm built on RFs that searches for stable, high-order interactions. Our method, the iterative random forest algorithm (iRF), sequentially grows feature-weighted RFs to perform soft dimension reduction of the feature space and stabilize decision paths. We decode the fitted RFs using a generalization of the random intersection trees algorithm (RIT) (20). This procedure

Significance

We developed a predictive, stable, and interpretable tool: the iterative random forest algorithm (iRF). iRF discovers high-order interactions among biomolecules with the same order of computational cost as random forests. We demonstrate the efficacy of iRF by finding known and promising interactions among biomolecules, of up to fifth and sixth order, in two data examples in transcriptional regulation and alternative splicing.

Author contributions: S.B., K.K., J.B.B., and B.Y. designed research, performed research, contributed analytic tools, analyzed data, and wrote the paper.

Reviewers: M.M.H., Princess Margaret Cancer Center; and D.J., Oak Ridge National Laboratory.

The authors declare no conflict of interest.

See Commentary on page 1690.

[1]S.B. and K.K. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: binyu@stat.berkeley.edu or jbbrown@lbl.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1711236115/-/DCSupplemental.

identifies high-order feature combinations that are prevalent on the RF decision paths. In addition to the high predictive accuracy of RFs, the decision tree base learner captures the underlying biology of local, combinatorial interactions (21), an important feature for biological data, where a single molecule often performs many roles in various cellular contexts. Moreover, invariance of decision trees to monotone transformations (12) to a large extent mitigates normalization issues that are a major concern in the analysis of genomics data, where signal-to-noise ratios vary widely even between biological replicates (22, 23). Using empirical and numerical examples, we show that iRF is competitive with RF in terms of predictive accuracy and extracts both known and compelling candidate interactions in two motivating biological problems in epigenomics and transcriptomics. An open-source R implementation of iRF is available through CRAN (https://cran.r-project.org/web/packages/iRF/index.html).

Our Method: Iterative RFs

The iRF algorithm searches for high-order feature interactions in three steps. First, iterative feature reweighting adaptively regularizes RF fitting. Second, decision rules extracted from a feature-weighted RF map from continuous or categorical to binary features. This mapping allows us to identify prevalent interactions in the RF through a generalization of the RIT, a computationally efficient algorithm that searches for high-order interactions in binary data (20). Finally, a bagging step assesses the stability of recovered interactions with respect to the bootstrap perturbation of the data. We briefly review the feature-weighted RF and RIT before presenting iRF.

Preliminaries: Feature-Weighted RF and RIT. To reduce the dimensionality of the feature space without removing marginally unimportant features that may participate in high-order interactions, we use a feature-weighted version of RF. Specifically, for a set of nonnegative weights $w = (w_1, \ldots, w_p)$, where p is the number of features, let $RF(w)$ denote a feature-weighted RF constructed with w. In $RF(w)$, instead of taking a uniform random sample of features at each split, one chooses the jth feature with probability proportional to w_j. Weighted-tree ensembles have been proposed in ref. 24 under the name "enriched random forests" and used for feature selection in genomic data analysis. Note that with this notation, Breiman's original RF amounts to $RF(1/p, \ldots, 1/p)$.

iRF builds upon a generalization of the RIT, an algorithm that performs a randomized search for high-order interactions among binary features in a deterministic setting. More precisely, the RIT searches for co-occurring collections of s binary features, or order-s interactions, that appear with greater frequency in a given class. The algorithm recovers such interactions with high probability (relative to the randomness it introduces) at a substantially lower computational cost than $O(p^s)$, provided the interaction pattern is sufficiently prevalent in the data and individual features are sparse. We briefly present the basic RIT algorithm and refer readers to the original paper (20) for a complete description.

Consider a binary classification problem with n observations and p binary features. Suppose we are given data in the form $(\mathcal{I}_i, Z_i)$, $i = 1, \ldots, n$. Here, each $Z_i \in \{0, 1\}$ is a binary label and $\mathcal{I}_i \subseteq \{1, 2, \ldots, p\}$ is a feature-index subset indicating the indexes of "active" features associated with observation i. In the context of gene transcription, $\mathcal{I}_i$ can be thought of as a collection of TFs and histone modifications with abnormally high or low enrichments near the ith gene's promoter region, and Z_i can indicate whether gene i is transcribed or not. With these notations, prevalence of an interaction $S \subseteq \{1, \ldots, p\}$ in the class $C \in \{0, 1\}$ is defined as

$$\mathbb{P}_n(S|Z = C) := \frac{\sum_{i=1}^n \mathbb{1}(S \subseteq \mathcal{I}_i)}{\sum_{i=1}^n \mathbb{1}(Z_i = C)},$$

where $\mathbb{P}_n$ denotes the empirical probability distribution and $\mathbb{1}(\cdot)$ the indicator function. For given thresholds $0 \le \theta_0 < \theta_1 \le 1$, the RIT performs a randomized search for interactions S satisfying

$$\mathbb{P}_n(S|Z = 1) \ge \theta_1, \quad \mathbb{P}_n(S|Z = 0) \le \theta_0. \qquad [1]$$

For each class $C \in \{0, 1\}$ and a prespecified integer D, let $j_1, ..., j_D$ be randomly chosen indexes from the set of observations $\{i : Z_i = C\}$. To search for interactions S satisfying condition **1**, the RIT takes D-fold intersections $\mathcal{I}_{j_1} \cap \mathcal{I}_{j_2} \cap \ldots \cap \mathcal{I}_{j_D}$ from the randomly selected observations in class C. To reduce computational complexity, these interactions are performed in a tree-like fashion (*SI Appendix*, section S1, Algorithm 1), where each nonleaf node has n_{child} children. This process is repeated M times for a given class C, resulting in a collection of survived interactions $\mathcal{S} = \bigcup_{m=1}^M \mathcal{S}_m$, where each $\mathcal{S}_m$ is the set of interactions that remains following the D-fold intersection process in tree $m = 1, \ldots, M$. The prevalences of interactions across different classes are subsequently compared using condition **1**. The main intuition is that if an interaction S is highly prevalent in a particular class, it will survive the D-fold intersection with high probability.

iRFs. The iRF algorithm places interaction discovery in a supervised learning framework to identify class-specific, active index sets required for the RIT. This framing allows us to recover high-order interactions that are associated with accurate prediction in feature-weighted RFs.

We consider the binary classification setting with training data $\mathcal{D}$ in the form $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, with continuous or categorical features $\mathbf{x} = (x_1, \ldots, x_p)$, and a binary label $y \in \{0, 1\}$. Our goal is to find subsets $S \subseteq \{1, \ldots, p\}$ of features, or interactions, that are both highly prevalent within a class $C \in \{0, 1\}$ and that provide good differentiation between the two classes. To encourage generalizability of our results, we search for interactions in ensembles of decision trees fitted on bootstrap samples of $\mathcal{D}$. This allows us to identify interactions that are robust to small perturbations in the data. Before describing iRF, we present a generalized RIT that uses any RF, weighted or not, to generate active index sets from continuous or categorical features. Our generalized RIT is independent of the other iRF components in the sense that other approaches could be used to generate the input for the RIT. We remark on our particular choices in *SI Appendix*, section S2.

Generalized RIT (Through an RF). For each tree $t = 1, \ldots, T$ in the output tree ensemble of an RF, we collect all leaf nodes and index them by $j_t = 1, ..., J(t)$. Each feature–response pair $(\mathbf{x}_i, y_i)$ is represented with respect to a tree t by $(\mathcal{I}_{i_t}, Z_{i_t})$, where $\mathcal{I}_{i_t}$ is the set of unique feature indexes falling on the path of the leaf node containing $(\mathbf{x}_i, y_i)$ in the tth tree. Hence, each $(\mathbf{x}_i, y_i)$ produces T such index set and label pairs, corresponding to the T trees. We aggregate these pairs across observations and trees as

$$\mathcal{R} = \{(\mathcal{I}_{i_t}, Z_{i_t}) : \mathbf{x}_i \text{ falls in leaf node } i_t \text{ of tree } t\} \qquad [2]$$

and apply RIT on this transformed dataset $\mathcal{R}$ to obtain a set of interactions.

We now describe the three components of iRF. A depiction is shown in Fig. 1 and the complete workflow is presented in *SI Appendix*, section S1, Algorithm 2. We remark on the algorithm further in *SI Appendix*, section S2.

1) Iteratively reweighted RF. Given an iteration number K, iRF iteratively grows K feature-weighted RFs $RF(w^{(k)})$, $k = 1, \ldots, K$, on the data $\mathcal{D}$. The first iteration of iRF ($k = 1$) starts with $w^{(1)} := (1/p, \ldots, 1/p)$ and stores the importance

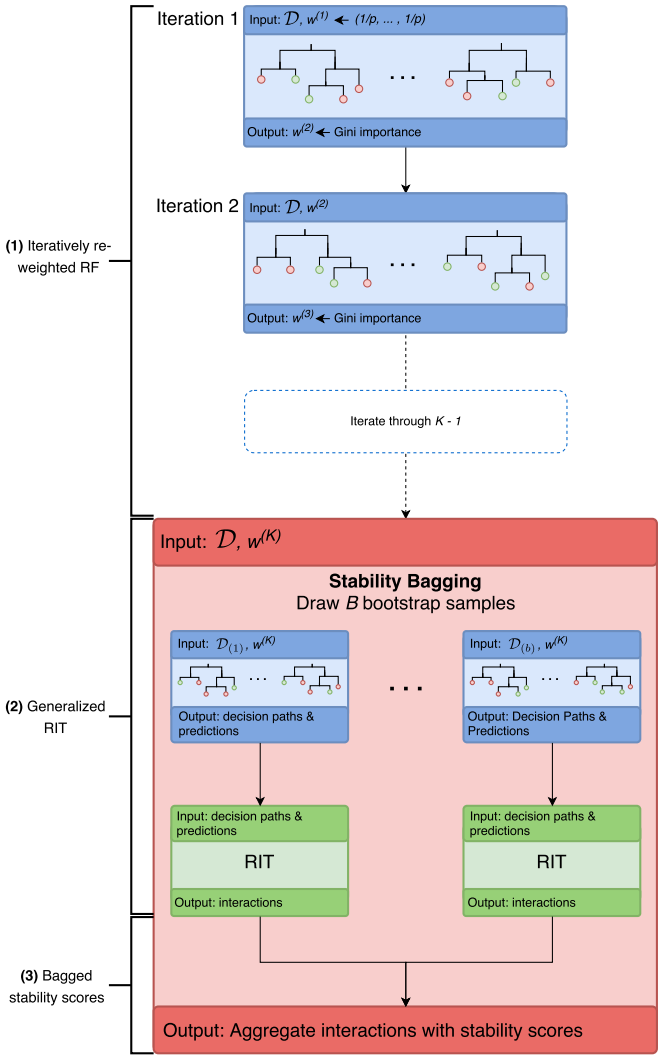

Fig. 1. iRF workflow. Iteratively reweighted RFs (blue boxes) are trained on full data $\mathcal{D}$ and pass Gini importance as weights to the next iteration. In iteration K (red box), feature-weighted RFs are grown using $w^{(K)}$ on B bootstrap samples of the full data $\mathcal{D}_{(1)}, \ldots, \mathcal{D}_{(B)}$. Decision paths and predicted leaf node labels are passed to the RIT (green box), which computes prevalent interactions in the RF ensemble. Recovered interactions are scored for stability across (outer-layer) bootstrap samples.

(mean decrease in Gini impurity) of the p features as $v^{(1)} = (v_1^{(1)}, \ldots, v_p^{(1)})$. For iterations $k = 2, \ldots, K$, we set $w^{(k)} = v^{(k-1)}$ and grow a weighted RF with weights set equal to the RF feature importance from the previous iteration. Iterative approaches for fitting RFs have been previously proposed in ref. 25 and combined with hard thresholding to select features in microarray data.

2) Generalized RIT (through RF($w^{(K)}$)). We apply the generalized RIT to the last feature-weighted RF grown in iteration K. That is, decision rules generated in the process of fitting $RF(w^{(K)})$ provide the mapping from continuous or categorical to binary features required for the RIT. This process produces a collection of interactions $\mathcal{S}$.

3) Bagged stability scores. In addition to bootstrap sampling in the weighted RF, we use an "outer layer" of bootstrapping to assess the stability of recovered interactions. We generate bootstrap samples of the data $\mathcal{D}_{(b)}, b = 1, \ldots, B$, fit $RF(w^{(K)})$ on each bootstrap sample $\mathcal{D}_{(b)}$, and use the generalized RIT to identify interactions $\mathcal{S}_{(b)}$ across each bootstrap sample. We define the stability score of an interaction $S \in \cup_{b=1}^{B} \mathcal{S}_{(b)}$ as

$$sta(S) = \frac{1}{B} \cdot \sum_{b=1}^{B} \mathbb{1}\{S \in \mathcal{S}_{(b)}\},$$

representing the proportion of times (out of B bootstrap samples) an interaction appears as an output of the RIT. This averaging step is exactly the bagging idea of Breimain (26).

iRF Tuning Parameters. The iRF algorithm inherits tuning parameters from its two base algorithms, RF and RIT. The predictive performance of RF is known to be highly resistant to choice of parameters (13), so we use the default parameters in the R randomForest package. Specifically, we set the number of trees `ntree` $= 500$ and the number of variables sampled at each node `mtry` $= \sqrt{p}$ and grow trees to purity. For the RIT algorithm, we use the basic version or algorithm 1 of ref. 20 and grow $M = 500$ intersection trees of depth $D = 5$ with $n_{child} = 2$, which empirically leads to a good balance between computation time and quality of recovered interactions. We find that both prediction accuracy and interaction recovery of iRF are fairly robust to these parameter choices (*SI Appendix*, section S2.6).

In addition to the tuning parameters of RF and RIT, the iRF workflow introduces two additional tuning parameters: (*i*) number of bootstrap samples B and (*ii*) number of iterations K. Larger values of B provide a more precise description of the uncertainty associated with each interaction at the expense of increased computation cost. In our simulations and case studies we set $B \in (10, 30)$ and find that results are qualitatively similar in this range. The number of iterations controls the degree of regularization on the fitted RF. We find that the quality of recovered interactions can improve dramatically for $K > 1$ (*SI Appendix*, section S5). In *Case Study I: Enhancer Elements in Drosophila* and *Case Study II: Alternative Splicing in a Human-Derived Cell Line*, we report interactions with K selected by fivefold cross-validation.

Simulation Experiments

We developed and tested iRF through extensive simulation studies based on biologically inspired generative models using both synthetic and real data (*SI Appendix*, section S5). In particular, we generated responses using Boolean rules intended to reflect the stereospecific nature of interactions among biomolecules (27). In total, we considered seven generative models built from and (AND), or (OR), and exclusive OR (XOR) rules, with the number of observations and features ranging from 100 to 5,000 and 50 to 2,500, respectively. We introduced noise into our models both by randomly swapping response labels for up to 30% of observations and through RF-derived rules learned on held-out data.

We find that the predictive performance of iRF ($K > 1$) is generally comparable with that of RF ($K = 1$). However, iRF recovers the full data-generating rule, up to an order-8 interaction in our simulations, as the most stable interaction in many settings where RF rarely recovers interactions of order >2. The computational complexity of recovering these interactions is substantially lower than that of competing methods that search for interactions incrementally (*SI Appendix*, section S6 and Fig. S18).

Our experiments suggest that iterative reweighting encourages iRF to use a stable set of features on decision paths (*SI Appendix*, Fig. S9). Specifically, features that are identified as important in early iterations tend to be selected among the first several splits in later iterations (*SI Appendix*, Fig. S10). This allows iRF to generate partitions of the feature space where marginally unimportant, active features become conditionally important and thus more likely to be selected on decision paths. For a full description of simulations and results, see *SI Appendix*, section S5.

Case Study I: Enhancer Elements in *Drosophila*

Development and function in multicellular organisms rely on precisely regulated spatiotemporal gene expression. Enhancers play a critical role in this process by coordinating combinatorial TF binding, whose integrated activity leads to patterned gene expression during embryogenesis (28). In the early *Drosophila* embryo, a small cohort of ∼40 TFs drive patterning (for a review see ref. 29), providing a well-studied, simplified model system in which to investigate the relationship between TF binding and enhancer activities. Extensive work has resulted in genome-wide, quantitative maps of DNA occupancy for 23 TFs (30) and 13 histone modifications (6), as well as labels of enhancer status for 7,809 genomic sequences in blastoderm (stage 5) *Drosophila* embryos (1, 31). See *SI Appendix*, section S3 for descriptions of data collection and preprocessing.

To investigate the relationship between enhancers, TF binding, and chromatin state, we used iRF to predict enhancer status for each of the genomic sequences (3,912 training, 3,897 test). We achieved an area under the precision-recall curve (AUC-PR) on the held-out test data of 0.5 for $K = 5$ (Fig. 2*A*). This corresponds to a Matthews correlation coefficient (MCC) of 0.43 [positive predictive value (PPV) of 0.71] when predicted probabilities are thresholded to maximize MCC in the training data.

Fig. 2*B* reports stability scores of recovered interactions for $K = 5$. We note that the data analyzed are whole embryo and interactions found by iRF do not necessarily represent physical complexes. However, for the well-studied case of pairwise TF interactions, 80% of our findings with stability score >0.5 have been previously reported as physical (*SI Appendix*, section S4 and Table S1). For instance, interactions among gap proteins *Giant* (*Gt*), Krüppel (Kr), and Hunchback (Hb), some of the most well-characterized interactions in the early *Drosophila* embryo (32), are all highly stable [$sta(Gt\text{-Kr}) = 1.0$, $sta(Gt\text{-Hb}) = 0.93$, $sta(\text{Hb-Kr}) = 0.73$]. Physical evidence supporting high-order mechanisms is a frontier of experimental research and hence limited, but our excellent pairwise results give us hope that high-order interactions we identify as stable have a good chance of being confirmed by follow-up work.

iRF also identified several high-order interactions surrounding the early regulatory factor *Zelda* (*Zld*) [$sta(Zld\text{-}Gt\text{-}Twi) = 1.0$, $sta(Zld\text{-}Gt\text{-Kr}) = 0.7$]. *Zld* has been previously shown to play an essential role during the maternal–zygotic transition (33, 34), and there is evidence to suggest that *Zld* facilitates binding to regulatory elements (35). We find that *Zld* binding in isolation rarely drives enhancer activity, but in the presence of other TFs, particularly the anterior–posterior (AP) patterning factors *Gt* and Kr, it is highly likely to induce transcription. This generalizes the dependence of Bicoid-induced transcription on *Zld* binding to several of the AP factors (36) and is broadly consistent with the idea that *Zld* is potentiating, rather than an activating factor (35).

More broadly, response surfaces associated with stable high-order interactions indicate AND-like rules (Fig. 2*C*). In other words, the proportion of active enhancers is substantially higher for sequences where all TFs are sufficiently bound, compared with sequences where only some of the TFs exhibit high levels of occupancy. Fig. 2*C* demonstrates a putative third-order interaction found by iRF ($sta(\text{Kr-}Gt\text{-}Zld) = 0.7$). In Fig. 2*C*, *Left*, the *Gt-Zld* response surface is plotted using only sequences for which Kr occupancy is lower than the median Kr level, and the proportion of active enhancers is uniformly low (<10%). The response surface in Fig. 2*C*, *Right* is plotted using only sequences where Kr occupancy is higher than median Kr level and shows that the proportion of active elements is as high as 60% when both *Zld* and *Gt* are sufficiently bound. This points to an order-3 AND rule, where all three proteins are required for enhancer activation in a subset of sequences. In Fig. 2*D*, we show the subset of sequences that correspond to this AND rule (highlighted in red), using a superheat map (37), which juxtaposes two separately clustered heat maps corresponding to active and inactive elements. Note that the response surfaces are drawn using held-out test data to illustrate the generalizability of interactions detected by iRF. While overlapping patterns of TF binding have been previously reported (30), to the best of our knowledge this is the first report of an AND-like response surface for enhancer activation. Third-order interactions have been studied in only a handful of enhancer elements, most notably eve stripe 2 (for a review see ref. 38), and our results indicate that they are broadly important for the establishment of early zygotic transcription and therefore body patterning.

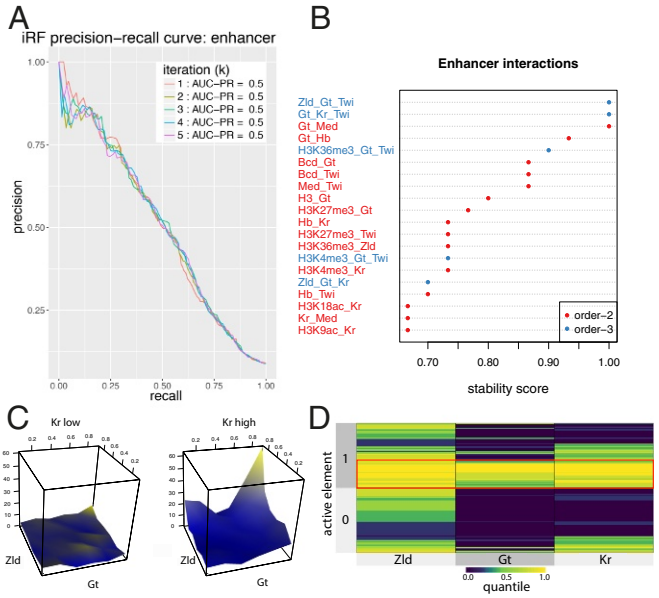

Fig. 2. (*A*) Accuracy of iRF (AUC-PR) in predicting active elements from TF binding and histone modification data. (*B*) The 20 most stable interactions recovered by iRF after five iterations. Interactions that are a strict subset of another interaction with stability score ≥0.5 have been removed for cleaner visualization. iRF recovers known interactions among *Gt*, Kr, and Hb and interacting roles of master regulator *Zld*. (*C*) Surface maps demonstrating the proportion of active enhancers by quantiles of *Zld*, *Gt*, and Kr binding (held-out test data). On the subset of data where Kr binding is lower than the median Kr level, the proportion of active enhancers does not change with *Gt* and *Zld*. On the subset of data with Kr binding above the median level, the structure of the response surface reflects an order-3 AND interaction: Increased levels of *Zld*, *Gt*, and Kr binding are indicative of enhancer status for a subset of observations. (*D*) Quantiles of *Zld*, *Gt*, and Kr binding grouped by enhancer status (balanced sample of held-out test data). The block of active elements highlighted in red represents the subset of observations for which the AND interaction is active.

Case Study II: Alternative Splicing in a Human-Derived Cell Line

In eukaryotes, alternative splicing of primary messenger RNA (mRNA) transcripts is a highly regulated process in which multiple distinct mRNAs are produced by the same gene. In the case of mRNAs, the result of this process is the diversification of the proteome and hence the library of functional molecules in cells. The activity of the spliceosome, the ribonucleoprotein responsible for most splicing in eukaryotic genomes, is driven by complex, cell-type–specific interactions with cohorts of RNA-binding proteins (RBP) (39, 40), suggesting that high-order interactions play an important role in the regulation of alternative splicing. However, our understanding of this system derives from decades of study in genetics, biochemistry, and structural biology.

Learning interactions directly from genomics data has the potential to accelerate our pace of discovery in the study of co- and posttranscriptional gene regulation.

Studies, initially in model organisms, have revealed that the chromatin mark H3K36me3, the DNA-binding protein CTCF, and a few other factors all play splice-enhancing roles (41–43). However, the extent to which chromatin state and DNA-binding factors interact *en masse* to modulate cotranscriptional splicing remains unknown (44). To identify interactions that form the basis of chromatin-mediated splicing, we used iRF to predict thresholded splicing rates for 23,823 exons [RNA-seq percent-spliced-in (PSI) values (https://github.com/guigolab/ipsa-nf); 11,911 training, 11,912 test], from ChIP-seq assays measuring enrichment of chromatin marks and TF-binding events (253 ChIP assays on 107 unique TFs and 11 histone modifications). Preprocessing methods are described in *SI Appendix*, section S3.

In this prediction problem, we achieved an AUC-PR on the held-out test data of 0.51 for $K = 2$ (Fig. 3*A*). This corresponds to a MCC of 0.47 (PPV 0.72) on held-out test data when predicted probabilities are thresholded to maximize MCC in the training data. Fig. 3*B* reports stability scores of recovered interactions for $K = 2$. We find interactions involving H3K36me3, a number of interactions involving other chromatin marks, and posttranslationally modified states of RNA Pol II. In particular, we find that the impact of serine 2 phosphorylation of Pol II appears highly dependent on local chromatin state. Remarkably, iRF identified an order-6 interaction surrounding H3K36me3 and S2 phospho-Pol II (stability score 0.5, Fig. 3 *B* and *C*) along with two highly stable order-5 subsets of this interaction (stability scores 1.0). A subset of highly spliced exons highlighted in red is enriched for all six of these elements, indicating a potential AND-type rule related to splicing events (Fig. 3*C*). This observation is consistent with, and offers a quantitative model for, the previously reported predominance of cotranscriptional splicing in this cell line (45). We note that the search space of order-6 interactions is $>10^{11}$ and that this interaction is discovered with an order-zero increase over the computational cost of finding important features using RF. Recovering such interactions without exponential speed penalties represents a substantial advantage over previous methods and positions our approach uniquely for the discovery of complex, nonlinear interactions.

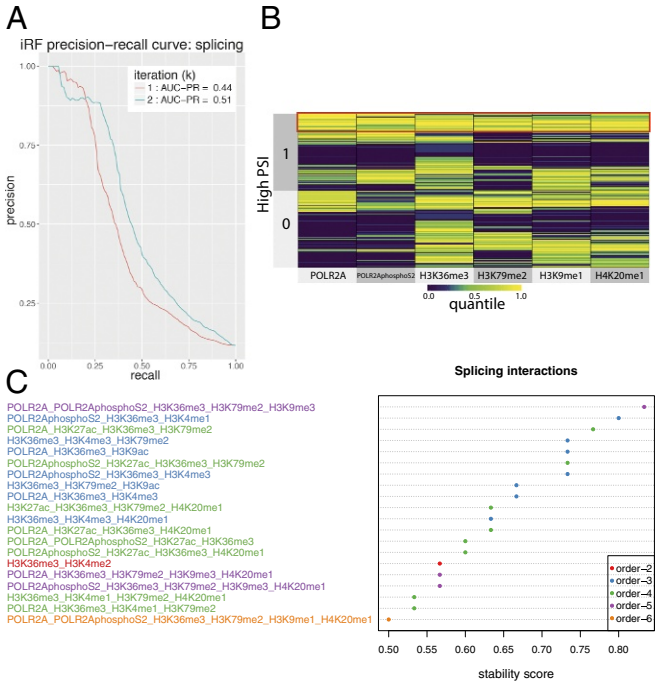

Fig. 3. (*A*) Accuracy of iRF (AUC-PR) in classifying included exons from excluded exons in held-out test data. iRF shows 7% increase in AUC-PR over RF. (*B*) An order-6 interaction recovered by iRF (stability score 0.5) displayed on a superheat map which juxtaposes two separately clustered heat maps of exons with high and low splicing rates. Coenrichment of all six plotted features reflects an AND-type rule indicative of high splicing rates for the exons highlighted in red (held-out test data). The subset of Pol II, S2 phospho-Pol II, H3K36me3, H3K79me2, and H4K20me1 was recovered as an order-5 interaction in all bootstrap samples (stability score 1.0). (*C*) The 20 most stable interactions recovered in the second iteration of iRF. Interactions that are a strict subset of another interaction with stability score $\geq$0.5 have been removed for cleaner visualization.

Discussion

Systems governed by nonlinear interactions are ubiquitous in biology. We developed a predictive and stable method, iRF, for learning such feature interactions. iRF identified known and promising interactions in early zygotic enhancer activation in the *Drosophila* embryo and posits more high-order interactions in splicing regulation for a human-derived system.

Validation and assessment of complex interactions in biological systems are necessary and challenging, but new wet-lab tools are becoming available for targeted genome and epigenome engineering. For instance, the CRISPR system has been adjusted for targeted manipulation of posttranslational modifications to histones (46). This may allow for tests to determine whether modifications to distinct residues at multivalent nucleosomes function in a nonadditive fashion in splicing regulation. Several of the histone marks that appear in the interactions we report, including H3K36me3 and H4K20me1, have been previously identified (47) as essential for establishing splicing patterns in the early embryo. Our findings point to direct interactions between these two distinct marks. This observation generates interesting questions: What proteins, if any, mediate these dependencies? What is the role of Phospho-S2 Pol II in the interaction? Proteomics on ChIP samples may help reveal the complete set of factors involved in these processes, and new assays such as Co-ChIP may enable the mapping of multiple histone marks at single-nucleosome resolution (48).

We have offered evidence that iRF constitutes a useful tool for generating hypotheses from the study of high-throughput genomics data, but many challenges await. iRF currently handles data heterogeneity only implicitly, and the order of detectable interaction depends directly on the depth of the tree, which is on the order of $\log_2(n)$. We are currently investigating local importance measures to explicitly relate discovered interactions to specific observations. This strategy has the potential to further localize feature selection and improve the interpretability of discovered rules. Additionally, iRF does not distinguish between interaction forms, for instance additive vs. nonadditive. We are exploring tests of rule structure to provide better insights into the precise form of rule–response relationships.

To date, machine learning has been driven largely by the need for accurate prediction. Leveraging machine-learning algorithms for scientific insights into the mechanics that underlie natural and artificial systems will require an understanding of why prediction is possible. The stability principle, which asserts that statistical results should at a minimum be reproducible across reasonable data and model perturbations, has been advocated in ref. 49 as a second consideration to work toward understanding and interpretability in science. Iterative and data-adaptive regularization procedures such as iRF are based on prediction and stability and have the potential to be widely adaptable to diverse algorithmic and computational architectures, improving interpretability and informativeness by increasing the stability of learners.

ACKNOWLEDGMENTS. We thank P. Bickel and S. Shrotriya for helpful comments, T. Arbel for preparing the *Drosophila* dataset, and S. Celniker for help in vetting the *Drosophila* data and for consultation on TF interactions. This research was supported in part by Grants National Human Genome Research Institute (NHGRI) U01HG007031, Army Research Office W911NF1710005, Office of Naval Research N00014-16-1-2664, Department of Energy (DOE) DE-AC02-05CH11231, NHGRI R00 HG006698, DOE (SBIR/STTR) Award DE-SC0017069, and National Science Foundation (NSF) DMS-1613002. We thank the Center for Science of Information, a US NSF Science and Technology Center, under Grant CCF-0939370. Research reported in this publication was supported by the National Library of Medicine of the NIH under Award T32LM012417. B.Y. acknowledges support from the Miller Institute for her Miller Professorship in 2016–2017. S.B. acknowledges the support of University of California, Berkeley, and Lawrence Berkeley National Laboratory, where he conducted work on this paper as a postdoc. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

1. Fisher WW, et al. (2012) DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proc Natl Acad Sci USA* 109:21330–21335.
2. Thomas S, et al. (2011) Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biol* 12:R43.
3. Li Xy, et al. (2008) Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* 6:e27.
4. Breeze CE, et al. (2016) eFORGE: A tool for identifying cell type-specific signal in epigenomic data. *Cell Rep* 17:2137–2150.
5. Hoffman MM, et al. (2012) Integrative annotation of chromatin elements from encode data. *Nucleic Acids Res* 41:827–841.
6. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.
7. Dong X, et al. (2012) Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol* 13:1–17.
8. Hota SK, Bruneau BG (2016) ATP-dependent chromatin remodeling during mammalian development. *Development* 143:2882–2897.
9. Zhou J, Troyanskaya OG (2014) Global quantitative modeling of chromatin factor interactions. *PLoS Comput Biol* 10:e1003525.
10. Lundberg SM, et al. (2016) Chromnet: Learning the human chromatin network from all ENCODE ChIP-seq data. *Genome Biol* 17:82.
11. Yoshida K, Yoshimoto J, Doya K (2017) Sparse kernel canonical correlation analysis for discovery of nonlinear interactions in high-dimensional data. *BMC Bioinformatics* 18:108.
12. Breiman L, Friedman J, Stone CJ, Olshen RA (1984) *Classification and Regression Trees* (CRC Press, Belmont, CA).
13. Breiman L (2001) Random forests. *Machine Learn* 45:5–32.
14. Meinshausen N (2010) Node harvest. *Ann Appl Stat* 4:2049–2072.
15. Meinshausen N (2009) Forest garrote. *Electron J Stat* 3:1288–1304.
16. Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. *Ann Appl Stat* 2:916–954.
17. Ruczinski CKI, LeBlanc ML, Hsu L (2001) Sequence analysis using logic regression. *Genet Epidemiol* 21:S626–S631.
18. Ritchie MD, et al. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 69:138–147.
19. Zhang Y, Liu JS (2007) Bayesian inference of epistatic interactions in case-control studies. *Nat Genet* 39:1167–1173.
20. Shah RD, Meinshausen N (2014) Random intersection trees. *J Machine Learn Res* 15:629–654.
21. Li G, et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148:84–98.
22. Landt SG, et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 22:1813–1831.
23. Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* 5:1752–1779.
24. Amaratunga D, Cabrera J, Lee YS (2008) Enriched random forests. *Bioinformatics* 24:2010–2014.
25. Anaissi A, Kennedy PJ, Goyal M, Catchpoole DR (2013) A balanced iterative random forest for gene selection from microarray data. *BMC Bioinformatics* 14:261.
26. Breiman L (1996) Bagging predictors. *Machine Learn* 24:123–140.
27. Nelson DL, Lehninger AL, Cox MM (2008) *Lehninger Principles of Biochemistry* (Macmillan, New York).
28. Levine M (2010) Transcriptional enhancers in animal development and evolution. *Curr Biol* 20:R754–R763.
29. Rivera-Pomar R, Jäckle H (1996) From gradients to stripes in Drosophila embryogenesis: Filling in the gaps. *Trends Genet* 12:478–483.
30. MacArthur S, et al. (2009) Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol* 10:R80.
31. Berman BP, et al. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. *Proc Natl Acad Sci USA* 99:757–762.
32. Nüsslein-Volhard C, Wieschaus E (1980) Mutations affecting segment number and polarity in Drosophila. *Nature* 287:795–801.
33. Liang HL, et al. (2008) The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature* 456:400–403.
34. Harrison MM, Li XY, Kaplan T, Botchan MR, Eisen MB (2011) Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS Genet* 7:e1002266.
35. Foo SM, et al. (2014) Zelda potentiates morphogen activity by increasing chromatin accessibility. *Curr Biol* 24:1341–1346.
36. Xu Z, et al. (2014) Impacts of the ubiquitous factor Zelda on Bicoid-dependent DNA binding and transcription in Drosophila. *Genes Dev* 28:608–621.
37. Barter RL, Yu B (2015) Superheat: Supervised heatmaps for visualizing complex data. arXiv:1512.01524.
38. Levine M (2013) Computing away the magic? *eLife* 2:e01135.
39. So BR, et al. (2016) A U1 snRNP-specific assembly pathway reveals the SMN complex as a versatile hub for RNP exchange. *Nat Struct Mol Biol* 23:225–230.
40. Stoiber MH, et al. (2015) Extensive cross-regulation of post-transcriptional regulatory networks in Drosophila. *Genome Res* 25:1692–1702.
41. Kolasinska-Zwierz P, et al. (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet* 41:376–381.
42. Sims Iii RJ, Reinberg D (2009) Processing the H3K36me3 signature. *Nat Genet* 41:270–271.
43. Kornblihtt AR (2012) CTCF: From insulators to alternative splicing regulation. *Cell Res* 22:450–452.
44. Allemand E, et al. (2016) A broad set of chromatin factors influences splicing. *PLoS Genet* 12:e1006318.
45. Tilgner H, et al. (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* 22:1616–1625.
46. Hilton IB, et al. (2015) Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat Biotechnol* 33: 510–517.
47. Hallmann RS, Schneeweiss LG, Correa E, Zamora J (1998) Fine needle aspiration biopsy of thymic carcinoid tumor: A case with immunocytochemical correlation. *Acta Cytol* 42:1042–1043.
48. Weiner A, et al. (2016) Co-ChIP enables genome-wide mapping of histone mark co-occurrence at single-molecule resolution. *Nat Biotechnol* 34:953–961.
49. Yu B (2013) Stability. *Bernoulli* 19:1484–1500.