frontiers in Genetics

METHODS
published: 30 August 2019
doi: 10.3389/fgene.2019.00729

Functional Site Discovery From Incomplete Training Data: A Case Study With Nucleic Acid–Binding Proteins

Wenchuan Wang[1†], Robert Langlois[2†], Marina Langlois[2], Georgi Z. Genchev[1,2,3], Xiaolei Wang[1,4] and Hui Lu[1,2,5]*

[1] SJTU-Yale Joint Center for Biostatistics and Data Science, Department of Bioinformatics and Biostatistics, College of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai, China, [2] Department of Bioengineering and Department of Computer Science, University of Illinois at Chicago, Chicago, IL, United States, [3] Bulgarian Institute for Genomics and Precision Medicine, Sofia, Bulgaria, [4] Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China, [5] Center for Biomedical Informatics, Shanghai Children's Hospital, Shanghai, China

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China

Reviewed by:
Weidong Tian,
Fudan University, China
Dong Xu,
University of Missouri, United States

***Correspondence:**
Hui Lu
huilu@sjtu.edu.cn

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 21 December 2018
Accepted: 11 July 2019
Published: 30 August 2019

Citation:
Wang W, Langlois R, Langlois M, Genchev GZ, Wang X and Lu H (2019) Functional Site Discovery From Incomplete Training Data: A Case Study With Nucleic Acid–Binding Proteins. Front. Genet. 10:729. doi: 10.3389/fgene.2019.00729

Function annotation efforts provide a foundation to our understanding of cellular processes and the functioning of the living cell. This motivates high-throughput computational methods to characterize new protein members of a particular function. Research work has focused on discriminative machine-learning methods, which promise to make efficient, *de novo* predictions of protein function. Furthermore, available function annotation exists predominantly for individual proteins rather than residues of which only a subset is necessary for the conveyance of a particular function. This limits discriminative approaches to predicting functions for which there is sufficient residue-level annotation, e.g., identification of DNA-binding proteins or where an excellent global representation can be divined. Complete understanding of the various functions of proteins requires discovery and functional annotation at the residue level. Herein, we cast this problem into the setting of multiple-instance learning, which only requires knowledge of the protein's function yet identifies functionally relevant residues and need not rely on homology. We developed a new multiple-instance leaning algorithm derived from AdaBoost and benchmarked this algorithm against two well-studied protein function prediction tasks: annotating proteins that bind DNA and RNA. This algorithm outperforms certain previous approaches in annotating protein function while identifying functionally relevant residues involved in binding both DNA and RNA, and on one protein-DNA benchmark, it achieves near perfect classification.

Keywords: machine learning, protein sequence and structural analysis, multiple-instance learning, decision trees, semi supervised learning, protein function annotation, DNA binding proteins, RNA binding proteins

INTRODUCTION

Computational tools have become indispensable in guiding, analyzing, and simulating the mechanistic details underlying experimental studies. Recent innovations in high-throughput experiments for function discovery have provided sufficient data to model and understand the characteristics that govern specific function using machine-learning methods. Such methods have been used to address biological problems ranging from microarray analysis and its application

in diagnosis, therapy decisions, and clinical testing (Juneau et al., 2014; Peterson et al., 2015; Shen et al., 2018); inter-disease relationships and similarities (Carson et al., 2017; Qin and Lu, 2018) image-based diagnostics (Mehta et al., 2017); predicting protein structural characteristics (Langlois and Lu, 2010a; Abbass and Nebel, 2015; Andreeva, 2016; Kashani-Amin et al., 2018) or clinically relevant discovery enabled by next-generation sequencing data of genomes and transcriptomes of diseased and normal cells (Gunaratne et al., 2012; Hayes and Kim, 2015; Gu et al., 2017; Liu et al., 2017; Gong et al., 2018; Liu et al., 2018).

High-throughput sequence and structural genomics projects have continued to outpace corresponding functional discovery projects producing a deluge of protein data, with only a fraction having some functional annotation. This annotation typically provides an indication of the general function but rarely, and when available—less reliably—provides mechanistic detail for a particular function. Systems biology research has focused on analyzing and predicting known interactions between proteins whereas pharmaceutical research requires greater knowledge in the mechanistic details of molecular function. Both efforts would benefit from machine-learning methods that can accurately classify protein function using the limited amount of training data available.

There are two approaches to the classification problem motivated by different statistical views: generative and discriminative learning. On one hand, the generative approach attempts to solve a more general problem i.e., modeling $[p(x,y)]$ providing greater flexibility at the cost of computational complexity. In order to design an efficient generative algorithm, strong assumptions must be made; e.g., in sequence alignment, one makes the assumption that sequence similarity equals function similarity. On the other hand, discriminative classifiers attempt to find a direct mapping between the class label (y) and the input vectors (x). Since this approach solves the specific problem at hand, rather than a more general problem, discriminative approaches should be preferred to generative ones (Libbrecht and Noble, 2015). However, the fact remains that generative, sequence alignment techniques remain predominant in the face of recently developed discriminative approaches. So, why have these discriminative techniques not been more successful? The fundamental problem seems to be that research has focused on a single type of discriminate method, classification, which requires labeled training examples. Since protein function annotation data is limited, only a few functional groups such as nucleic acid–binding proteins provide sufficient labeled training data.

A number of discriminative techniques have been developed to deal with incomplete knowledge of the training data such as: semi-supervised learning (Chapelle et al., 2010), active learning (Reker and Schneider, 2015), positive and unlabeled learning (Bhardwaj et al., 2010), and multiple-instance learning (MIL) (Carbonneau et al., 2018). While the first three approaches have demonstrated that unlabeled training data can be used to improve learning, the last approach leverages additional information, i.e., labeled groupings of unlabeled data. In MIL, examples (also referred to as instances) are organized into groups called bags. The class label is associated with the bag rather than the instance; the bag is labeled positive if at least one instance in the bag is labeled positive; otherwise, the bag is labeled negative. Consider the functional site discovery problem: functional data usually pertains to the protein rather than to specific functional sites. Hence, in the MIL formulation, the protein is a labeled bag and the residues (or motifs or pockets) are the instances belonging to that protein/bag.

MIL was originally developed for handwritten digit recognition by Keeler et al. (1990) and was later popularized by Dietterich et al. (1997) to predict drug activity. It has subsequently been applied to a number of problem domains including context-based image retrieval (Maron and Lozano-Perez, 1998; Andrews et al., 2003a), protein super-family annotation (TrX proteins) (Scott et al.), and text categorization (Ray and Craven, 2005). A number of algorithms have been developed to solve MIL including convolutional neural networks (Keeler et al., 1990), axis parallel (Dietterich et al., 1997), support-vector machines (Doran and Ray, 2014), diverse density (Maron and Lozano-Perez, 1998), and standard binary classifiers (Ray and Craven, 2005).

MIL algorithm–based approaches have recently found increased use in the diagnosis of cancer (Li et al., 2015; Mercan et al., 2018; Yousefi et al., 2018), application in neurology for classification of brain abnormalities (Tong et al., 2014), and the prediction of phenotype from metagenomics data (Rahman et al., 2017) to name a few. Recent work has utilized MIL-based methods to predict major histocompatibility complex class II (MHC-II)–binding peptides (Xu et al., 2014) and transcription factor-DNA interaction (Gao et al., 2015; Gao and Ruan, 2017).

The boosting framework has also been conscripted to solve MIL problems. These approaches fall into two groups: modify the weak learner or modify the boosting cost function. That is, Auer and Ortner (2004) took the first approach by boosting a weak MIL-algorithm based on hyper-balls. Other algorithms have been developed using the second approach. For example, Andrews and Hofmann (2003b) used disjunctive logic programming (Lee and Grossmann, 2000) to create a boosting algorithm that achieves a large margin for at least one instance in each bag. Likewise, other groups (Xu and Frank, 2004; Viola et al., 2006) have used a derivation of the AnyBoost framework (Mason et al., 1999) to design an MIL cost function, which can be solved by numerical optimization.

Our work herein formulates the function prediction problem in the setting of MIL. In our approach, the function of a protein is identified through the discovery of key residue microenvironments that strongly signal the existence of a particular functional site. This method requires only two sets of example sequences or structures: one that has the function of interest and another that does not. We do not require knowledge of the functional sites yet this method automatically discovers such sites in order to predict the function of the protein. In the formulation of this approach, we predict function rather than superfamily assignment of a protein; moreover, we represent the protein by each residue's microenvironment rather than by pre-calculated conserved motifs.

To solve this problem, we developed a novel boosting algorithm (Langlois, 2008) derived from the AdaBoost framework (Schapire and Singer, 1999) that efficiently and accurately identifies residue

microenvironments that correspond to functional sites. We then benchmark this approach on two protein function assignment problems: the identification of DNA- and RNA-binding proteins. These proteins play an essential role in nearly every cellular process. A number of experimental (Cajone et al., 1989; Freeman et al., 1995; Chou et al., 2003; Buck and Lieb, 2004; Nutiu et al., 2011; Gordan et al., 2013) and computational (Bhardwaj et al., 2005; Szilagyi and Skolnick, 2006; Bhardwaj and Lu, 2007; Langlois et al., 2007; Tjong and Zhou, 2007; Gao and Skolnick, 2009; Langlois and Lu, 2010b; Weirauch et al., 2013; Xu et al., 2015) approaches have been developed to identify these proteins and their functional sites. Since DNA- and RNA-binding proteins provide a substantial number of labeled examples, e.g., residues known to bind DNA or RNA, these problems have been studied extensively thus presenting an excellent proof of concept for our approach.

RESULTS

We demonstrate the ability of an MIL algorithm to accurately predict the function of a protein using its constituent residues with two benchmark nucleic-acid binding datasets: DNA- and RNA-binding proteins. The characteristics of each dataset are summarized in **Table 1**. Both datasets have been used in previous studies to identify residues that bind DNA (Szilagyi and Skolnick, 2006; Langlois et al., 2007) and RNA (Terribilini et al., 2006; Langlois et al., 2007; Kumar et al., 2008). During training, each residue in a DNA-binding protein is considered DNA-binding and in a non-DNA-binding protein non-binding during training and cross-validation. Nevertheless, these residue-level labels are used for later evaluation of the algorithm on the residue level.

Protein Function Annotation

We compare two learning algorithms to solve the MIL problem: AdaBoost and AdaBoost.C2MIL on decision trees. The first algorithm, AdaBoost on decision trees is a classification algorithm, which views MIL as a classification problem with positive class noise (Blum, 1998). While other classifiers have been extensively tested on MIL problems (Ray and Craven, 2005), AdaBoost on decision trees has not; this is due to its past poor performance on problems with mislabeled data (Schapire, 1999). The second algorithm AdaBoost.C2MIL is a modification of the original AdaBoost algorithm we developed specifically to handle MIL, which gives special treatment to instances (residues) in a positive bag (DNA-binding protein).

Table 2 summarizes the performance of each algorithm in terms of area under the receiver operating characteristic (ROC) curve on the protein-level (first column), residue-level over the entire dataset (second column), and over just the DNA-binding proteins (third column). The protein-level results demonstrate the effectiveness of the proposed C2MIL variant over the standard AdaBoost algorithm where C2MIL outperforms AdaBoost by 5% on the DNA-binding task and by 6% on the RNA-binding task. The residue-level performance over the entire dataset is worse in both cases. However, this is due to the inclusion of residues from non-binding proteins, which skew the results. When considering the more pertinent case of only nucleic acid–binding proteins, the C2MIL algorithm outperforms AdaBoost in both cases: by almost 9% for the DNA-binding task and 3% for the RNA-binding task.

The performance over the DNA-binding set on the protein-level exceeds several previously published works. First, the performance of the C2MIL algorithm achieves 95.8% area under the ROC whereas the best previous result was 93% (Szilagyi and Skolnick, 2006) and 91.0% (Langlois and Lu, 2010b). At 85.0% specificity, C2MIL achieves 94.4% sensitivity compared to 89.0% (Szilagyi and Skolnick, 2006). At 95.0% specificity, Stawiski et al. (Stawiski et al., 2003) achieved 81.0% sensitivity while C2MIL 86.1% sensitivity. Finally, at 98% specificity, Langlois and Lu (Langlois and Lu, 2010b) achieved 48.1% sensitivity and C2MIL 70.8% sensitivity. Overall, C2MIL shows marked improvement in accurately predicting whether a protein binds DNA.

Functional Site Prediction

Since no residue-level labels were given during training, i.e., the algorithm does not know which residues bind DNA or RNA,

TABLE 1 | Tabulates the number of proteins in both the DNA and RNA datasets.

	Total	Protein Positive	Negative	Total	Residue Positive	Negative
DNA	310	60	250	109,826	2,505	107,321
RNA	304	80	224	91,538	3,235	88,303

TABLE 2 | Performance of algorithms in the multiple-instance learning (MIL) function prediction task—area under the receiver operating characteristic (ROC) curve.

		Protein	Residue (All)	Residue (-NA)
DNA binding	AdaBoost	90.3	84.4	63.2
	AdaBoost.C2MIL	95.8	82.7	72.1
RNA binding	AdaBoost	84.1	79.4	65.6
	AdaBoost.C2MIL	90.2	74.5	68.7

the performance of C2MIL is significantly less than the current best: 72% (**Table 1**) *versus* 83% (Langlois et al., 2007) in terms of area under the ROC. At the same time, the performance over the full dataset (both DNA-binding and non-binding proteins) is significantly better than over just the DNA-binding proteins: 82.7% area under the ROC (**Table 1**). This seems to indicate that non-binding residue environments or substructures on non-NA-binding proteins are easier to predict than corresponding ones on NA-binding proteins.

The ROC plots in **Figure 1** and **Figure 2** compare the performance of C2MIL with the standard AdaBoost algorithm over the DNA-binding dataset. In **Figure 1A**, both algorithms cross several times with no clear winner. However, at low false-positive rates (**Figure 1B**), C2MIL dominates the standard AdaBoost providing an explanation for C2MIL's better performance on the protein level. Since only a single residue predicted positive means the entire bag is positive, this is the important region on the residue-level ROC curve.

The ROC plots in **Figure 2** compare the performance of C2MIL with the standard AdaBoost algorithm over the residues from only DNA-binding proteins. This evaluation follows that of other DNA-binding papers (Langlois et al., 2007). On this task, C2MIL dominates the standard AdaBoost algorithm over the entire range of the ROC plot. As the protein-level results indicate, C2MIL finds at least one residue microenvironment that strongly indicates a given protein is DNA binding. Moreover, these instance-level results demonstrate that not many residues fit the bill given the rather low sensitivity at low false-positive rates.

Trends in Residue-Level Prediction

To better understand the residue microenvironments that characterize NA-binding proteins, we plot each type of residue which has been correctly predicted DNA binding in terms of recall and precision (**Figure 3**). Precision measures the fraction of residues predicted NA binding that are actually DNA binding

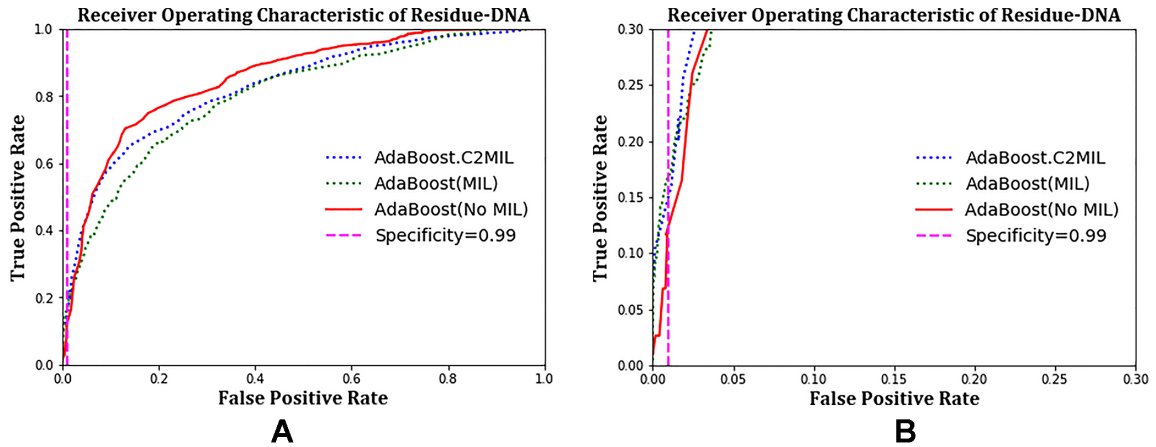

FIGURE 1 | Comparison of learning tasks and algorithms on the residue-level over the entire dataset using a receiver operating characteristic curve: **(A)** entire curve and **(B)** zoomed on the 99% specificity.

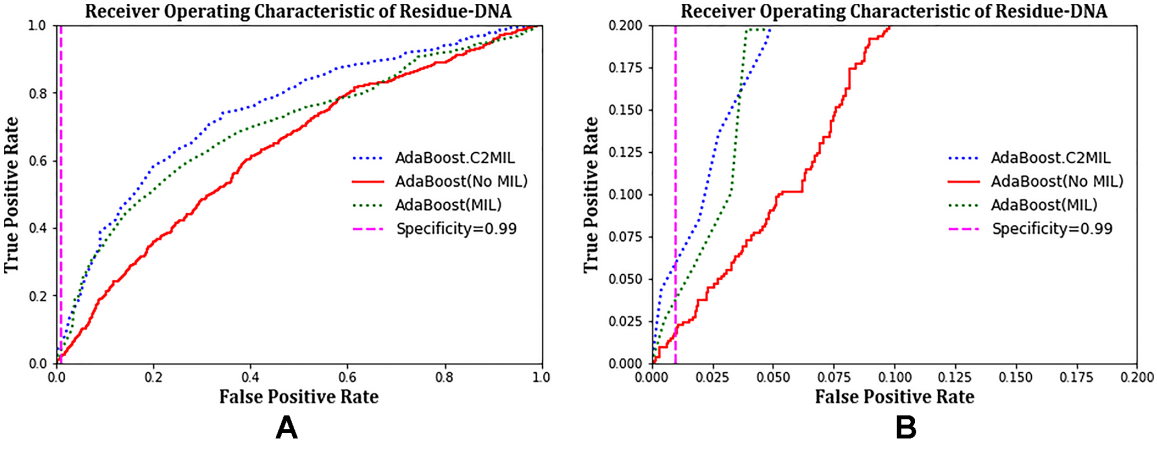

FIGURE 2 | Comparison of learning tasks and algorithms on the residue-level over only DNA-binding proteins using a receiver operating characteristic curve: **(A)** entire curve and **(B)** zoomed on the 99% specificity.

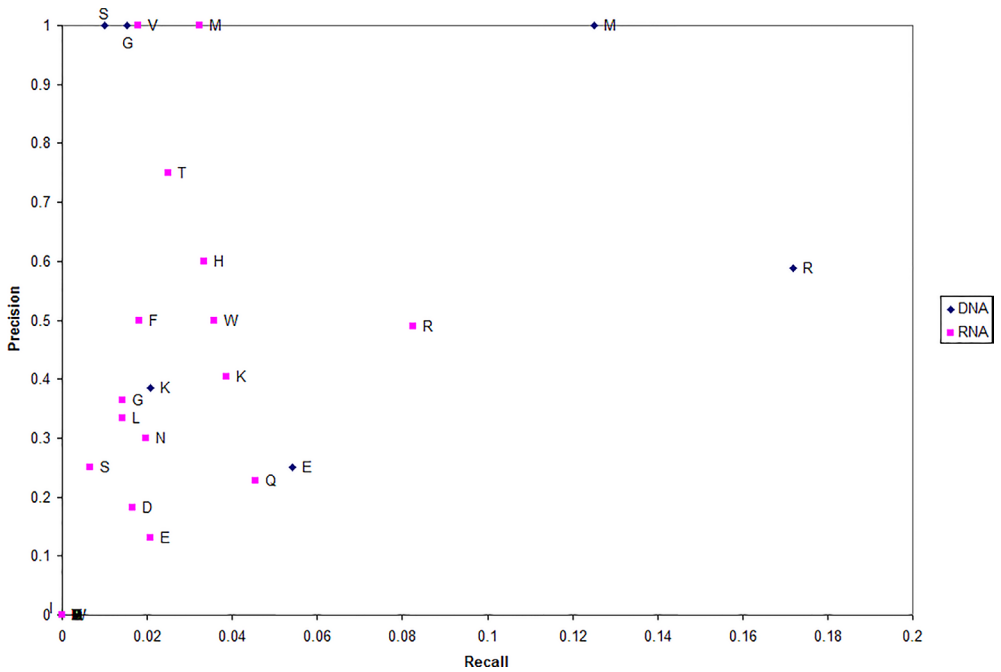

FIGURE 3 | Precision and recall for each nucleic acid (NA)-binding protein residue type.

(in blue) and RNA binding (in red). Recall measures the fraction of NA-binding residues correctly predicted NA binding.

The first trend evident in **Figure 3** is that far more residues can be used to predict a protein RNA binding (red) as opposed to DNA binding (blue). This suggests that more residues are involved in protein-RNA interactions than protein-DNA. Second, arginine is unsurprising the dominant residue predicted for both NA-binding proteins.

Third, DNA-binding proteins can unexpectedly be well characterized by microenvironments centered on either serine (S) or glycine (G) with a precision of 1.0; e.g., every serine predicted as DNA binding actually was DNA binding. While previous works have suggested glycine (specifically its content) as more correlated with the non-binding set (Bhardwaj et al., 2005; Szilagyi and Skolnick, 2006; Langlois et al., 2007), it has been observed that glycine can make non-specific interactions with DNA (Luscombe and Thornton, 2002) and that glycine-rich linkers are critical to regulatory protein function (Singh et al., 2014).

Fourth, a set of RNA-binding proteins can be accurately characterized by microenvironment centered on either valine (V) or methionine (M) with a precision of 1.0. These residues as well as histidine and threonine have been found important experimentally. Threonine has been shown to make specific interactions with both splice sites (Colwill et al., 1996; Zhang and Fuller, 2003) and rRNA (Clemens et al., 1993). Likewise, histidine has been found important for specificity (Hake et al., 1998) and valine makes unique interactions with viral RNA (Pinck et al., 1970).

Note that, in proteins predicted DNA/RNA binding, these four residues (V, M, S, and G) provide a rough location the NA-binding site each protein. This demonstrates that the MIL-algorithm identifies DNA-/RNA-binding proteins based on residue important to their function.

DISCUSSION

Conventional approaches that apply machine learning to function prediction have relied on a global representation of the sequence or structure, or a local representation of a residue's environment on a target protein. In the first case, only examples of known proteins with a particular function are required whereas the second case requires knowing the location of the active sites. Our proposed approach is similar to sequence alignment techniques in that we require only knowing the function of a particular protein and not the functional residues. Moreover, similar to sequence analysis techniques, it identifies a subset of probable functional residues. Nevertheless, our proposed algorithm does not require sequence similarity or homology to be effective (unlike sequence analysis techniques).

In this work, we demonstrate the ability of our MIL algorithm–based approach to identify potential binding sites and, through the presence of such a site, the function of the protein. This is done without knowledge of the binding sites during the training process. Essentially, one can both identify the function of and locate a binding site on a test protein without knowing, during the training process, the location of such sites. One can view MIL over structure-based features as sub-structure analysis where were consider a sliding window along the amino acid chain throughout the structure. Thus, a user only requires knowledge of the protein function, not the particular site, yet the resulting learning algorithm can predict both.

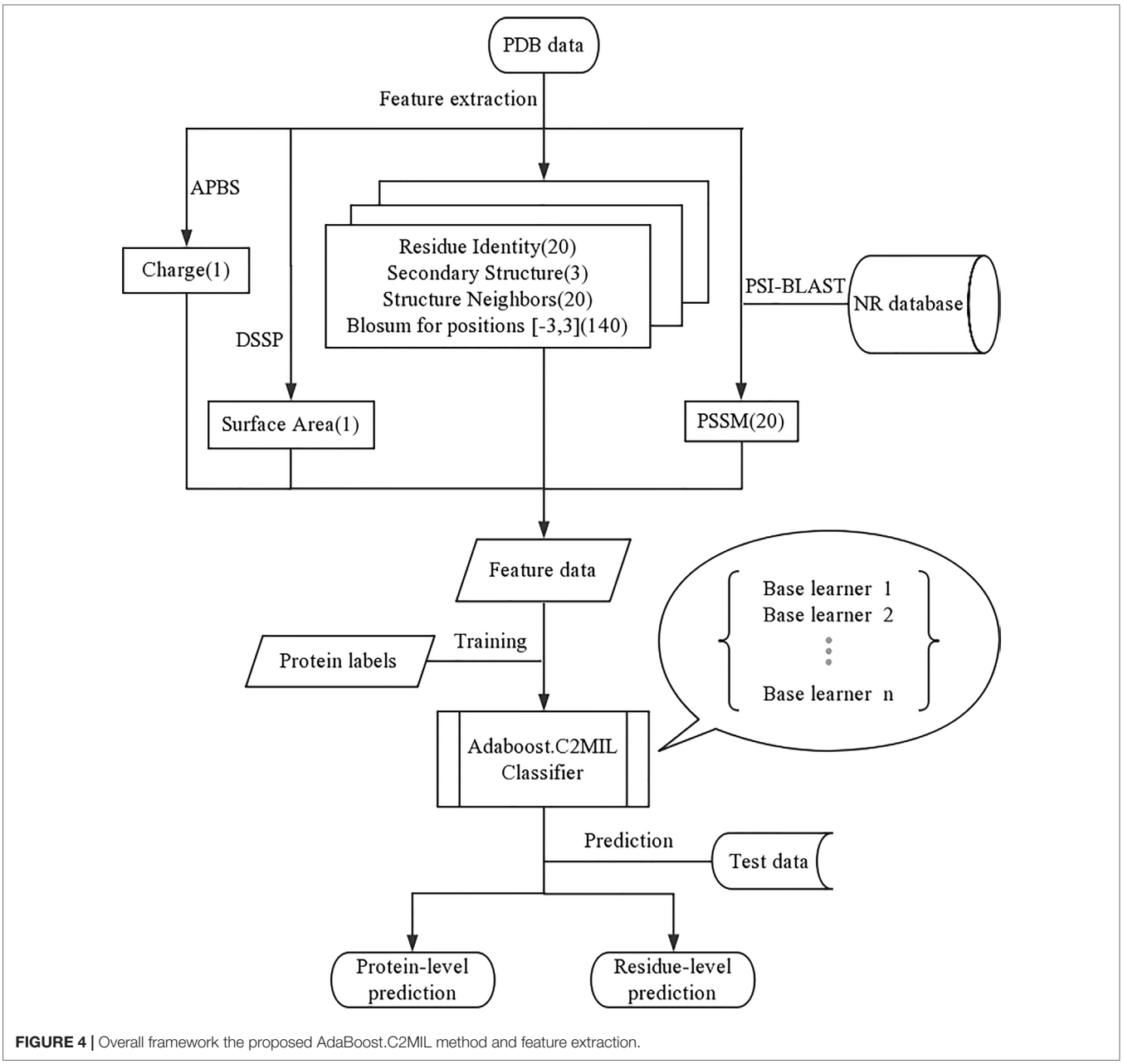

FIGURE 4 | Overall framework the proposed AdaBoost.C2MIL method and feature extraction.

The proposed approach also has several advantages over traditional homology-based methods:

- Does not rely on finding a similar structure/sequence
- Discovers functional sites with little prior knowledge

Our method does not require homologous sequences or structures; instead, it relies on physio-chemical characteristics in combination with (when available) structural features. It can also be applied to problems where knowledge of the functional site is limited. We also provide an analysis of MIL algorithms on the instance level. In some previously published MIL works, the authors evaluate their algorithms on the bag-level since instance-level labels are either unavailable or unreasonably expensive to obtain.

This works establishes the ability of our MIL algorithm–based method to outperform classification in discriminating RNA- or DNA-binding proteins from non-binding proteins. The success of this approach relies on the better representation of function permitted by the MIL problem formulation. Instead of representing the protein sequence or structure by some global representation, the MIL approach allows the entire protein to be decomposed into potential functional units and discovers which unit actually performs the function. Developing a feature encoding for a single functional unit is far easier than for the entire protein sequence or structure.

While multiple-instance (MI) learning has several advantages over classification, it remains a harder learning problem in that the learning algorithm does not have access to instance-level labels.

Nevertheless, the experiments clearly show that the proposed MI learner does not perform substantially worse when identifying residues that bind DNA or RNA. Indeed, these results compare favorably with the current state-of-the-art in residue classification.

There are several limitations to the present work. First, we do not limit the algorithm to only sequence information; yet, this will provide the primary source of data for this application. Second, this work does not consider open conformations, e.g., proteins not in complex with DNA. Since the current set of features does not require the exact residue orientation, this may not be a significant limitation. Third, it does not incorporate known binding residues; such residues can provide more information regarding these residues. This problem can be remedied through the application of active MIL (Zhang et al., 2008). Fourth, this algorithm would utilize and would benefit from far larger datasets such as sequences in the UniProt (Leinonen et al., 2004) database. Finally, the analysis of the important residues was just a first-order approximation to the potential wealth of information this technique can glean from both sequence and structural data.

MATERIALS AND METHODS

Dataset

There are two stringent benchmark datasets used for DNA- and RNA-binding protein prediction tasks. The first set is 60 DNA-binding proteins and 250 non-DNA-binding proteins derived by Liu et al. (2014) and later used by Shen et al. (2017) and Wei et al. (2017) (**Supplementary Table 1**). The second set is 80 RNA-binding proteins and 224 non-RNA-binding proteins used by Miao and Westhof (Miao and Westhof, 2015) and Paz et al. (2016) (**Supplementary Table 2**). The two datasets are both acquired from the Protein Data Bank, and short sequences (less than 50 amino acids) and sequences containing the consecutive character "X" have been removed. To eliminate the redundancy and homology bias that likely leads to overestimated performance, it removes sequences with ≥25% pairwise sequence identity to any other sequences in the dataset using the program CD-HIT.

Each residue in the protein is represented using the following features (feature count within parenthesis) (**Figure 4**):

- Residue identity (Bhardwaj et al., 2010)
- Secondary structure (Shen et al., 2018)
- Structure neighbors (Bhardwaj et al., 2010)
- PSSM for residue at that position (Bhardwaj et al., 2010)
- BLOSUM for positions $-3…3$ (140)
- Properties: Charge, Surface Area (Juneau et al., 2014)

The residue identifier is a 20-dimensional vector where the residue type is indicated by a non-zero value in the corresponding column. Likewise, there is a corresponding secondary structure identifier feature vector. The structure neighbors count the frequency of each residue type within 3 Å (measured heavy atom to heavy atom). The PSSM feature scores the conservation of this residue position. The BLOSUM window also estimates the residue conservation within a window around the specific residue. Finally, the properties of charge and surface area are estimated for each residue. For more details concerning the feature representation, see Langlois et al. (2007).

Algorithm

The Adaptive Boosting (AdaBoost) algorithm transforms a weak classifier $L(\cdot)$ into a strong ensemble classifier $H(\cdot)$(44). AdaBoost proves most effective with decision trees as the weak classifier (often referred to as "the best off-the-shelf classifier") and has one tunable parameter: the number of boosting iterations (T). Rather than the general boosting framework as in prior work (Mason et al., 1999), we propose to modify the AdaBoost algorithm itself to reduce MIL to importance-weighted classification.

EQUATION 1 | Proposed AdaBoost.C2MIL Algorithm

Given: $\{(X_1, y_1) \ldots (X_n, y_n)\}$
where : $X_i = \left\{\vec{x}_1 \ldots \vec{x}_{n_i}\right\}$ and $y_i \in \{-1, 1\}$ and $\vec{x}_j \in X$
Reorganize dataset such each negative bag contains one instance

Initialize: $w_{t=1,i} = \frac{1}{n}, i = 1 \ldots n$

For $t = 1 \ldots T$

1. Map dataset to instance level: $\hat{D} = \left(\vec{x}_{i,j}, y_i, \frac{w_i}{n_i}\right)$
2. Train weak classifier on instance-level dataset $L(\hat{D})$
3. Get confidence-rated instance-level hypothesis $\hat{h}_t : X \to \Re$
4. $\hat{p} = \frac{1}{1 + exp(-\hat{h}_t)}$
5. Get weak bag-level hypothesis: $h_t(X_i) = \frac{\Sigma_j \hat{p}_t(\vec{x}_{i,j}) sgn\left[\hat{h}_t\left(\vec{x}_{i,j}\right)\right]}{\Sigma_j \hat{p}_t(\vec{x}_{i,j})}$
6. $\varepsilon_t = \sum_{sgn[h_t(X_i)] \neq y_i} w_{t,i}$
7. $\alpha_t = \frac{1}{2} ln\left(\frac{1 - \varepsilon_t}{\varepsilon_t}\right)$
8. $w_{t+1,i} = w_{t,i} \cdot exp(-\alpha_t sgn[h_t(X_i)] sgn[y_i]) / Z_t, i = 1 \ldots n$

Output:

$$H(\vec{x}_{i,j}) = \sum_1^T \alpha_t \hat{h}_t(\vec{x}_{i,j}) \quad \text{Instance-level prediction}$$

$$H(X_i) = \sum_1^T \alpha_t h_t(X_i) \quad \text{Bag-level prediction}$$

The proposed algorithm, AdaBoost.C2MIL, is outlined in **Equation 1**. The first step in the algorithm is to set up the dataset. It starts by reorganizing the dataset such that each negative instance becomes its own bag while the positive instances remain grouped in their original bags. Note that, since we know each instance in a negative bag must be negative, this step does not disregard useful information. It then sets up a uniform weighted distribution on the bag level. Since each negative instance is a bag, it has its own weight whereas instances in a positive bag share a single weight.

The second step, within the for loop, starts by mapping the MIL dataset to a classification dataset where every instance in a positive bag is labeled positive, and the weight is split uniformly among the instances. Next, the algorithm trains a weak classifier (L) over the current distribution of the dataset, which gives confidence-rated hypothesis $\hat{h}_t$. The confidence-rated prediction follows (Schapire and Singer, 1999) and can be converted to a probability using the sigmoid function. Finally, for positive bags (and negative bags during evaluation), the bag-level prediction is a summation of the instance-level predictions (step 5).

The rest of the algorithm follows AdaBoost on the bag level. First, the algorithm estimates the bag-level error and then calculates the step size α. This step size is then used to increase the weight on incorrectly predicted bags and decrease on correctly predicted.

The output of the ensemble multiple-instance learner acts on both the bag and instance level. Each classifier contributes to the prediction of an instance whereas the bag-level prediction is made by the equation in step 5.

Experiments

The overall framework of our experiment is represented in **Figure 4**. The AdaBoost algorithm requires a weak learner and, as a weak learner, the decision tree works well across the board; we use a custom implementation with a top-down (Kearns and Mansour, 1996) impurity function for confidence-rated boosting. The algorithms, metrics, and graphs used in this work were generated using python. The performance is measured using 5-fold stratified cross-validation. And code is available at https://github.com/WintrumWang/AdaBoost.C2MIL.

AUTHOR CONTRIBUTIONS

RL and HL designed the project, WW and RL performed the project, ML and GG helped in method development and manuscript writing, XW participated in the computation. All authors approved the writing.

FUNDING

This work is partially supported by National Key R&D Program of China 2018YFC0910500, the Neil Shen's SJTU Medical Research Fund, SJTU-Yale Collaborative Research Seed Fund; NSFC 31728012, Science and Technology Commission of Shanghai Municipality (STCSM) grant 17DZ 22512000, Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01), LCNBI and ZJLab. RL acknowledges the support from NIH training grant T32 HL 07692.

ACKNOWLEDGMENTS

We thank Matthew B. Carson for assist in the dataset preparation and Irina Irodova for useful discussions regarding the proposed C2MIL algorithm.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00729/full#supplementary-material

SUPPLEMENTARY TABLE 1 | The DNA-binding protein benchmark dataset.

SUPPLEMENTARY TABLE 2 | The RNA-binding protein benchmark dataset.

REFERENCES

Abbass, J., and Nebel, J. C. (2015). Customised fragments libraries for protein structure prediction based on structural class annotations. *BMC Bioinform.* 16, 136. doi: 10.1186/s12859-015-0576-2

Andreeva, A. (2016). Lessons from making the Structural Classification of Proteins (SCOP) and their implications for protein structure modelling. *Biochem. Soc. Trans.* 44 (3), 937–943. doi: 10.1042/BST20160053

Andrews S., and Hofmann T., editors. (2003b). "Multiple instance learning via disjunctive programming boosting," in *Advances in Neural Information Processing Systems* (Vancouver, Whistler, Canada: MIT Press).

Andrews S, Tsochantaridis I, Hofmann T, editors. (2003a). "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems* (Vancouver, Whistler, Canada: MIT Press).

Auer P., and Ortner R., editors. A boosting approach to multiple instance learning. *European Conference on Machine Learning, Machine Learning: ECML 2004, Proceedings*, Pisa, Italy. (2004) 3201, 63–74.

Bhardwaj, N., and Lu, H. (2007). Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. *FEBS Lett.* 581 (5), 1058–1066. doi: 10.1016/j.febslet.2007.01.086

Bhardwaj, N., Gerstein, M., and Lu, H. (2010). Genome-wide sequence-based prediction of peripheral proteins using a novel semi-supervised learning technique. *BMC Bioinform.* 11 (1), S6. doi: 10.1186/1471-2105-11-S1-S6

Bhardwaj, N., Langlois, R. E., Zhao, G. J., and Lu, H. (2005). Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res.* 33 (20), 6486–6493. doi: 10.1093/nar/gki949

Blum, A. (1998). Kalai A. A note on learning from multiple-instance examples. *Mach. Learn.* 30 (1), 23–29. doi: 10.1023/A:1007402410823

Buck, M. J., and Lieb, J. D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* 83 (3), 349–360. doi: 10.1016/j.ygeno.2003.11.004

Cajone, F., Salina, M., and Benellizazzera, A. (1989). 4-Hydroxynonenal induces a DNA-binding protein similar to the heat-shock factor. *Biochem. J.* 262 (3), 977–979. doi: 10.1042/bj2620977

Carbonneau, M.-A., Cheplygina, V., Granger, E., and Gagnon, G. (2018). Multiple instance learning: a survey of problem characteristics and applications. *Pattern Recognit.* 77, 329–353. doi: 10.1016/j.patcog.2017.10.009

Carson, M. B., Liu, C., Lu, Y., Jia, C., and Lu, H. (2017). A disease similarity matrix based on the uniqueness of shared genes. *BMC Med. Genomics* 10 (Suppl 1), 26. doi: 10.1186/s12920-017-0265-2

Chapelle, O., Schölkopf, B., and Zien, A., (2010). *Semi-supervised learning*. 1st MIT Press pbk. ed. Cambridge, Mass.: MIT Press. x, 508 p.

Chou, C. C., Lin, T. W., Chen, C. Y., and Wang, A. H. J. (2003). Crystal structure of the hyperthermophilic archaeal DNA-binding protein Sso10b2 at a resolution of 1.85 angstroms. *J. Bacteriol.* 185 (14), 4066–4073. doi: 10.1128/JB.185.14.4066-4073.2003

Clemens, K., Wolf, V., McBryant, S., Zhang, P., Liao, X., Wright, P., et al. (1993). Molecular basis for specific recognition of both RNA and DNA by a zinc finger protein. *Science* 260 (5107), 530–533. doi: 10.1126/science.8475383

Colwill, K., Pawson, T., Andrews, B., Prasad, J., Manley, J. L., Bell, J. C., et al. (1996). The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intranuclear distribution. *EMBO J.* 15 (2), 265–275. doi: 10.1002/j.1460-2075.1996.tb00357.x

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* 89 (1–2), 31–71. doi: 10.1016/S0004-3702(96)00034-3

Doran, G., and Ray, S. (2014). A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. *Mach. Learn.* 97 (1), 79–102. doi: 10.1007/s10994-013-5429-5

Freeman, K., Gwadz, M., and Shore, D. (1995). Molecular and genetic-analysis of the toxic effect of rap1 overexpression in yeast. *Genetics* 141 (4), 1253–1262.

Gao, M., and Skolnick, J. (2009). From nonspecific DNA-protein encounter complexes to the prediction of DNA-protein interactions. *PLoS Comput. Biol.* 5 (4) 1–12. doi: 10.1371/journal.pcbi.1000341

Gao, Z., and Ruan, J. (2015). A structure-based multiple-instance learning approach to predicting *in vitro* transcription factor-DNA interaction. *BMC Genomics* 16 Suppl 4, S3. doi: 10.1186/1471-2164-16-S4-S3

Gao, Z., and Ruan, J. (2017). Computational modeling of *in vivo* and *in vitro* protein-DNA interactions by multiple instance learning. *Bioinformatics* 33 (14), 2097–2105. doi: 10.1093/bioinformatics/btx115

Gong, X., Jiang, J., Duan, Z., and Lu, H. (2018). A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology. *BMC Bioinform.* 19 (Suppl 4), 162. doi: 10.1186/s12859-018-2064-y

Gordan, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., et al. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.* 3 (4), 1093–1104. doi: 10.1016/j.celrep.2013.03.014

Gu, J. L., Chukhman, M., Lu, Y., Liu, C., Liu, S. Y., and Lu, H. (2017). RNA-seq based transcription characterization of fusion breakpoints as a potential estimator for its oncogenic potential. *Biomed. Res. Int.* 2017, 9829175. doi: 10.1155/2017/9829175

Gunaratne, P. H., Coarfa, C., Soibam, B., and Tandon, A. (2012). miRNA data analysis: next-gen sequencing. *Methods Mol. Biol.* 822, 273–288. doi: 10.1007/978-1-61779-427-8_19

Hake, L. E., Mendez, R., and Richter, J. D. (1998). Specificity of RNA binding by CPEB: requirement for RNA recognition motifs and a novel zinc finger. *Mol. Cell. Biol.* 18 (2), 685–693. doi: 10.1128/MCB.18.2.685

Hayes, D. N., and Kim, W. Y. (2015). The next steps in next-gen sequencing of cancer genomes. *J. Clin. Invest.* 125 (2), 462–468. doi: 10.1172/JCI68339

Juneau, K., Bogard, P. E., Huang, S., Mohseni, M., Wang, E. T., Ryvkin, P., et al. (2014). Microarray-based cell-free DNA analysis improves noninvasive prenatal testing. *Fetal. Diagn. Ther.* 36 (4), 282–286. doi: 10.1159/000367626

Kashani-Amin, E., Tabatabaei-Malazy, O., Sakhteman, A., Larijani, B., and Ebrahim-Habibi, A. (2019). A systematic review on popularity, application and characteristics of protein secondary structure prediction tools. *Curr. Drug Discov. Technol* 16(2), 159–172. doi: 10.2174/1570163815666180227162157

Kearns M., and Mansour Y., editors. On the boosting ability of top-down decision tree learning algorithms. *ACM Symposium on the Theory of Computing, Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing,* Philadelphia, Pennsylvania: ACM. (1996) 459–468.

Keeler J.D., Rumelhart D.E., Leow W.-K., editors. (1990). "Integrated segmentation and recognition of hand-printed numerals," in *Advances in Neural Information Processing Systems* (Denver, Colorado: Morgan Kaufmann Publisher).

Kumar, M., Gromiha, M. M., and Raghava, G. P. (2008). Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins* 71 (1), 189–194. doi: 10.1002/prot.21677

Langlois. (2008). *Machine Learning in Bioinformatics: Algorithms, Implementations and Applications.* Chicago: University of Illinois at Chicago.

Langlois, R. E., and Lu, H. (2010a). Machine learning for protein structure and function prediction. *Ann. Rep. Comp. Chem.* 4, 41–66. doi: 10.1016/S1574-1400(08)00003-0

Langlois, R. E., and Lu, H. (2010b). Boosting the prediction and understanding of DNA-binding domains from sequence. *Nucleic Acids Res.* 38 (10), 3149–3158. doi: 10.1093/nar/gkq061

Langlois, R. E., Carson, M. B., Bhardwaj, N., and Lu, H. (2007). Learning to translate sequence and structure to function: identifying DNA binding and membrane binding proteins. *Ann. Biomed. Eng.* 35 (6), 1043–1052. doi: 10.1007/s10439-007-9312-z

Lee, S., and Grossmann., I. E. (2000). New algorithms for nonlinear generalized disjunctive programming. *Comput. Chem. Eng. J.* 24 (9-10), 2125–2141. doi: 10.1016/S0098-1354(00)00581-0

Leinonen, R., Diez, F. G., Binns, D., Fleischmann, W., Lopez, R., and Apweiler, R. (2004). UniProt archive. *Bioinformatics* 20 (17), 3236–3237. doi: 10.1093/bioinformatics/bth191

Li, C., Shi, C., Zhang, H., Chen, Y., and Zhang, S. (2015). Multiple instance learning for computer aided detection and diagnosis of gastric cancer with dual-energy CT imaging. *J. Biomed. Inform.* 57, 358–368. doi: 10.1016/j.jbi.2015.08.017

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16 (6), 321–332. doi: 10.1038/nrg3920

Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X., et al. (2014). iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PLoS One* 9 (9), e106691. doi: 10.1371/journal.pone.0106691

Liu, C., Wang, X., Genchev, G. Z., and Lu, H. (2017). Multi-omics facilitated variable selection in cox-regression model for cancer prognosis prediction. *Methods* 124, 100–107. doi: 10.1016/j.ymeth.2017.06.010

Liu, S. Y., Wang, X. J., Qin, W. Y., Genchev, G. Z., and Lu, H. (2018). Transcription factors contribute to differential expression in cellular pathways in lung adenocarcinoma and lung squamous cell carcinoma. *Interdiscip. Sci.* 10 (4), 836–847. doi: 10.1007/s12539-018-0300-9

Luscombe, N. M., and Thornton, J. M. (2002). Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.* 320 (5), 991–1009. doi: 10.1016/S0022-2836(02)00571-5

Maron O., and Lozano-Perez T., editors. (1998). "A framework for multiple-instance learning," in *Advances in Neural Information Processing Systems* (Denver, Colorado: MIT Press).

Mason L., Baxter J., Bartlett P., and Frean M., editors. (1999). "Boosting algorithms as gradient descent," in *Advances in Neural Information Processing Systems* (Denver, Colorado: MIT Press).

Mehta, R., Cai, K., Kumar, N., Knuttinen, M. G., Anderson, T. M., Lu, H., et al. (2017). A lesion-based response prediction model using pretherapy PET/CT image features for Y90 radioembolization to hepatic malignancies. Technol. *Cancer Res. Treat.* 16 (5), 620–629. doi: 10.1177/1533034616666721

Mercan, C., Aksoy, S., Mercan, E., Shapiro, L. G., Weaver, D. L., and Elmore, J. G. (2018). Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. *IEEE Trans. Med. Imaging* 37 (1), 316–325. doi: 10.1109/TMI.2017.2758580

Miao, Z. C., and Westhof, E. (2015). Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. *Nucleic Acids Res.* 43 (11), 5340–5351. doi: 10.1093/nar/gkv446

Nutiu, R., Friedman, R. C., Luo, S., Khrebtukova, I., Silva, D., Li, R., et al. (2011). Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.* 29 (7), 659–664. doi: 10.1038/nbt.1882

Paz, I., Kligun, E., Bengad, B., and Mandel-Gutfreund, Y. (2016). BindUP: a web server for non-homology-based prediction of DNA and RNA binding proteins. *Nucleic Acids Res.* 44 (W1), W568–W574. doi: 10.1093/nar/gkw454

Peterson, J. F., Aggarwal, N., Smith, C. A., Gollin, S. M., Surti, U., Rajkovic, A., et al. (2015). Integration of microarray analysis into the clinical diagnosis of hematological malignancies: how much can we improve cytogenetic testing? *Oncotarget* 6 (22), 18845–18862. doi: 10.18632/oncotarget.4586

Pinck, M., Yot, P., Chapeville, F., and Duranton, H. M. (1970). Enzymatic binding of valine to the 3' end of TYMV-RNA. *Nature* 226 (5249), 954–956. doi: 10.1038/226954a0

Qin, W., and Lu, H. (2018). A novel joint analysis framework improves identification of differentially expressed genes in cross disease transcriptomic analysis. *BioData Min.* 11, 3. doi: 10.1186/s13040-018-0163-y

Rahman, M. A., LaPierre, N., and Rangwala, H. (2017). Phenotype prediction from metagenomic data using clustering and assembly with multiple instance learning (CAMIL). *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1–1 doi: 10.1109/TCBB.2017.2758782

Ray, S., and Craven, M., editors. Supervised versus multiple instance learning: an empirical comparison. *International Conference on Machine Learning, Proceedings of the 22nd International Conference on Machine Learning,* Bonn, Germany: ACM. (2005) 697–704.

Reker, D., and Schneider, G. (2015). Active-learning strategies in computer-assisted drug discovery. *Drug Discov. Today* 20 (4), 458–465. doi: 10.1016/j.drudis.2014.12.004

Schapire, R. E. Theoretical views of boosting and applications. *Proceedings of the 10th International Conference on Algorithmic Learning Theory*, *Algorithmic Learning Theory, Proceedings,* 735768: Springer-Verlag. (1999) 1720, p. 13–25.

Schapire, R. E., and Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.* 37 (3), 297–336. doi: 10.1023/A:1007614523901

Scott, S. D., Ji, H., Wen, P., Fomenko, D. E., and Gladyshev, V. N. On modeling protein superfamilies with low primary sequence conservation. Technical report. University of Nebraska, 2003 UNL-CSE-2003-4.

Shen, C., Ding, Y. J., Tang, J. J., Song, J., and Guo, F. (2017). Identification of DNA-protein binding sites through multi-scale local average blocks on sequence information. *Molecules* 22 (12). doi: 10.3390/molecules22122079

Shen, Y., Zhang, J., Fu, Z., Zhang, B., Chen, M., Ling, X., et al. (2018). Gene microarray analysis of the circular RNAs expression profile in human gastric cancer. *Oncol. Lett.* 15 (6), 9965–9972. doi: 10.3892/ol.2018.8590

Singh, N. S., Kachhap, S., Singh, R., Mishra, R. C., Singh, B., and Raychaudhuri, S. (2014). The length of glycine-rich linker in DNA-binding domain is critical for optimal functioning of quorum-sensing master regulatory protein HapR. *Mol. Genet. Genomics* 289 (6), 1171–1182. doi: 10.1007/s00438-014-0878-5

Stawiski, E. W., Gregoret, L. M., and Mandel-Gutfreund, Y. (2003). Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.* 326 (4), 1065–1079. doi: 10.1016/S0022-2836(03)00031-7

Szilagyi, A., and Skolnick, J. (2006). Efficient prediction of nucleic acid binding function from low-resolution protein structures. *J. Mol. Biol.* 358 (3), 922–933. doi: 10.1016/j.jmb.2006.02.053

Terribilini, M., Lee, J. H., Yan, C., Jernigan, R. L., Honavar, V., and Dobbs, D. (2006). Prediction of RNA binding sites in proteins from amino acid sequence. *RNA* 12 (8), 1450–1462. doi: 10.1261/rna.2197306

Tjong, H., and Zhou, H. X. (2007). DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res.* 35 (5), 1465–1477. doi: 10.1093/nar/gkm008

Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J. V., Rueckert, D., et al. (2014). Multiple instance learning for classification of dementia in brain MRI. *Med. Image Anal.* 18 (5), 808–818. doi: 10.1016/j.media.2014.04.006

Viola, P., Platt, J. C., and Zhang, C., (2006). "Multiple instance boosting for object detection," in *Advances in Neural Information Processing Systems*, vol. 18. Eds. Y. Weiss, B. Scholkopf, and J. C. Platt (Sudbury, Massachusetts, USA: MIT Press).

Wei, L. Y., Tang, J. J., and Zou, Q. (2017). Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inf. Sci.* 384, 135–144. doi: 10.1016/j.ins.2016.06.026

Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., et al. (2013). Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* 31 (2), 126–134. doi: 10.1038/nbt.2486

Xu, X., and Frank, E. (2004). "Logistic regression and boosting for labeled bags of instances," in *Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science*, vol. 3056. (Heidelberg: Springer), 272–281. doi: 10.1007/978-3-540-24775-3_35

Xu, R., Zhou, J., Wang, H., He, Y., Wang, X., and Liu, B. (2015). Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC Syst. Biol.* 9 Suppl 1, S10. doi: 10.1186/1752-0509-9-S1-S10

Xu, Y., Luo, C., Qian, M., Huang, X., and Zhu, S. (2014). MHC2MIL: a novel multiple instance learning based method for MHC-II peptide binding prediction by considering peptide flanking region and residue positions. *BMC Genomics* 15 Suppl 9, S9. doi: 10.1186/1471-2164-15-S9-S9

Yousefi, M., Krzyzak, A., and Suen, C. Y. (2018). Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning. *Comput. Biol. Med.* 96, 283–293. doi: 10.1016/j.compbiomed.2018.04.004

Zhang, W., and Fuller, G. N. (2003). *Genomic and Molecular Neuro-Oncology*. Vancouver, British Columbia, Canada: Jones & Bartlett Publishers.

Zhang, D., Wang, F., Shi, Z. W., and Zhang, C. S. (2008). Localized content based image retrieval by multiple instance active learning. *IEEE Image Proc.*, 921–924. doi: 10.1109/ICIP.2008.4711906

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.