54–70 *Nucleic Acids Research, 2018, Vol. 46, No. 1* *Published online 25 November 2017*
doi: 10.1093/nar/gkx1166

Integrating sequence and gene expression information predicts genome-wide DNA-binding proteins and suggests a cooperative mechanism

Shandar Ahmad[1,2,*], Philip Prathipati[2], Lokesh P. Tripathi[2], Yi-An Chen[2], Ajay Arya[1], Yoichi Murakami[2] and Kenji Mizuguchi[2]

[1]School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India and [2]Laboratory of Bioinformatics, National Institutes of Biomedical Innovation, Health and Nutrition, 7-6-8 Saito-asagi, Ibaraki, Osaka 5670085, Japan

Received December 03, 2016; Revised October 29, 2017; Editorial Decision November 05, 2017; Accepted November 15, 2017

ABSTRACT

DNA-binding proteins (DBPs) perform diverse biological functions ranging from transcription to pathogen sensing. Machine learning methods can not only identify DBPs *de novo* but also provide insights into their DNA-recognition dynamics. However, it remains unclear whether available methods that can accurately predict DNA-binding sites in *known* DBPs can also identify novel DBPs. Moreover, sequence information is blind to the cellular- and disease-specific contexts of DBP activities, whereas the under-utilized knowledge from public gene expression data offers great promise. To address these issues, we have developed novel methods for predicting DBPs by integrating sequence and gene expression-derived features and applied them to explore human, mouse and *Arabidopsis* proteomes. While our sequence-based models outperformed the gene expression-based ones, some proteins with weaker DBP-like sequence features were correctly predicted by gene expression-based features, suggesting that these proteins acquire a tangible DBP functionality in a conducive gene expression environment. Analysis of motif enrichment among the co-expressed genes of top 100 candidates DBPs from hitherto unannotated genes provides further avenues to explore their functional associations.

INTRODUCTION

DNA-binding proteins (DBPs) perform diverse biological functions ranging from pathogen recognition, transcription initiation and regulation and DNA packaging; they are also involved in modifications such as methylation and acetylation (1–14). Despite their functional diversity, however, they share remarkably similar attributes such as biases in the overall and binding site-local amino acid compositions. This feature allows a relatively accurate identification of DBPs from sequence or structural information alone without necessitating further characterization (15,16).

In general, the DNA-binding site residues (DBS) of DBPs are enriched in positively charged Arg residues, a signal which is further fine-tuned by their sequence and structural environments (17). These compositional biases can be accurately captured by statistical and machine learning models trained over carefully prepared non-redundant and accurately characterized datasets of DNA-binding proteins (18–20). These datasets are almost always derived from the known three-dimensional structures of protein–DNA complexes and do not include any non-DBPs (21,22). Thus, these trained models represent an internal discrimination of the DBS from the rest of the amino acid sequence and it is unclear whether they can also distinguish DBPs from other proteins. DBP prediction models, on the other hand, exploit the compositional biases in the DBPs compared to other proteins and these biases are not exactly the same as the DBS biases (16,23).

While a number of methods have been proposed for predicting DBPs and DBS separately (15,16,20,21,23–38), to the best of our knowledge, no study has been conducted to develop a prediction system that employs DBS as an engine for the DBP prediction, combined with the amino acid compositional biases of the full length proteins, and to evaluate it comprehensively on an entire genome.

In this study, we first investigated the various levels of DBP annotations, ranging from the existence of protein–DNA complexes in the crystal structures to protein domain assignments (39) and gene ontology (GO) term associations. For each level of DBP annotations, we examined the enrichment of features derived from the predicted DBS and provided background scores to these predictions. To ensure a strong predictive performance, we also predicted binding

*To whom correspondence should be addressed. Tel: +91 11 2674 8788; Fax: +91 11 2674 1586; Email: shandar@jnu.ac.in

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

residues for adenosine triphosphate (ATP), carbohydrates, RNA and proteins using our previously published methods (40–43). This step was performed to exclude the binding sites for other ligands from the prediction models as the sequence descriptors for different types of binding sites are very similar and may prove to be a confounding factor. To these scores, we added the whole protein amino acid composition and trained models for the entire human proteome using these features. This procedure resulted in a highly accurate and elaborately benchmarked method for DBP prediction. Top scoring novel predictions were manually examined to assess their potential for being DBPs.

Next, we evaluated an alternative approach to DBP prediction via global expression analysis of their source genes. Gene expression (GE) profiles and the features derived from them are promising for two reasons. First, it may be possible to annotate DBPs directly from the expression profiles of their coding genes in the same way as the prediction of more general gene functions (44–46). Such GE-based annotations would be especially useful if the sequences alone were insufficient to confer a DBP function on a gene (e.g. if the function was altered by the GE context). Global GE profiles have been previously employed for gene function prediction based on the principle of *'guilt by association'*, whereby a gene is assigned a particular function if the other genes co-expressed with it are known to have that function (44,45,47). Here, we go beyond the *guilt by association* principle and specifically investigate the distributions of expression levels (ELs), degree of co-expression with other genes and populations of GO terms in the co-expression network. Rather than assigning a function directly by *guilt-by-association*, we employ a machine learning approach to take into account the enriched and depleted GO terms with appropriate and implicit weights assigned to each occurrence of the GO terms.

Finally, as a natural corollary to the two analyses above, we investigated the interplay between the GE and sequence information in determining DBP annotations. Specifically, we investigated if the GE profiles could improve the accuracy of the sequence-based DBP prediction and whether specific functional subgroups showed varied levels of improvement. Further characterization of the top 100 currently unannotated, prioritized candidate DBPs in the human proteome, suggested that two of them, C6ORF23 and C6ORF15 are likely to be novel transcription factors (TFs).

Although the current study was performed primarily on human proteins, the analytical methods described here are of a general nature and can easily be extended to other systems, as we have demonstrated here using mouse and *Arabidopsis* proteins.

MATERIALS AND METHODS

The experimental design of the current study is summarized in Figure 1.

The first step involves developing methods to predict DBPs or DBP-coding genes independently from their sequences and GE patterns.

Subsequently, the two independently generated scores are integrated to gain insights into their interplay in predicting if a protein binds to its DNA targets. Individual steps illustrated in Figure 1 are explained below:

Predicting DBPs from sequence features

Genome-wide proteins datasets and DBP annotations. This study is focussed on the prediction of human DBPs, even though the methods developed are general and easily applicable to other datasets, as we demonstrate on two other organisms. Detailed procedures to compile sequence datasets from the human proteome and their DNA-binding annotations are explained in Supplementary Methods SM1–3. A brief summary is provided here.

UniProtKB is the primary source of sequence and annotation data used in this study (48). At the time of carrying out this study, UniProt consisted of 20,195 manually curated human proteins (the SwissProt dataset). This dataset will be referred to as SP_human in the rest of the manuscript. The term DNA-binding often differs in scope, as used by different sources of annotation. In the strictest sense of these annotations, a direct interaction between some residues from the protein and the target DNA is required, while a broad-based annotation may be assigned to proteins that participate in a protein–DNA interaction but indirectly, e.g. as a co-factor of a TF. To create a working database of both these groups of annotations, we relied upon four DNA-binding annotation sources: (i) the 'Sequence features' field of Uniprot entries (hereafter referred to as SeqFT), (ii) occurrences of sequence domains assigned a DNA-binding annotation in the Pfam database, (iii) sequence similarity with proteins that were observed in protein–DNA complex structures in the Protein Data Bank (PDB) and (iv) GO annotations. Proteins annotated as DNA-binding in each of these repositories were labelled "DBP bySeqFT", "DBP byPfam", "DBP byPDB" and "DBP byGO", respectively and the number of DBPs in each

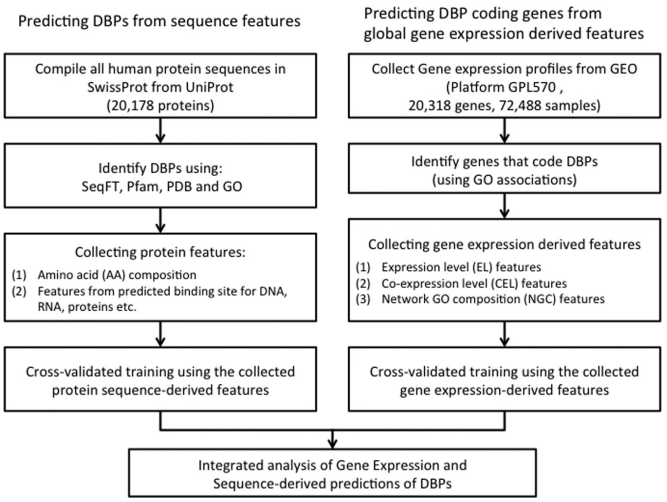

Figure 1. Overall experimental design of the study: DNA-binding protein predictions were performed independently using the sequence-derived data and the gene-expression profile-derived data in multiple steps and using different ways for feature extraction. The optimum models obtained through these twin steps were then compared in terms of performance and insights into the biological mechanism.

of these categories were 601, 684, 361 and 2407, respectively. Overlaps between these groups and a discussion of key observations from this analysis are provided in Supplementary Methods SM3.

Computing sequence features of proteins for analysis and prediction. We have previously developed methods for predicting DNA-binding site residues in proteins from single sequences as well as from evolutionary profiles using neural networks and other machine learning models (49,50). Predictive features in each case share essentially the same local sequence features, even though class labels and hence the trained models differ depending on the carefully prepared datasets (24–27,51–54). These methods have been recently reviewed comprehensively in the context of their predictive performances (28,51).

Many of these prediction models are based on a direct or indirect assessment of conservation scores derived from evolutionary profiles. Interestingly, methods for detecting binding residues for different ligands such as RNA, ATP and carbohydrates also rely on evolutionary features and it is therefore, likely that high scores for binding residues for one ligand type can be obtained for proteins that may actually bind to a different ligand type. Thus, a simultaneous prediction of binding residues for several ligands can provide more useful information about detecting DBPs *de novo* than from predicted DBS alone.

We have previously reported that even though DNA-binding sites are enriched in positively charged residues (primarily arginine), DBPs as such may not have identical biases. Indeed, DNA-binding can be achieved by the internal compositional biases and concentration of positively charged residues in one region of the structure, leading to the dipole-based recognition of DNA (55). Since distant charged residues can come together to form such regions, whole-protein level compositional biases are difficult to detect. On the other hand, there may be signals in the DBPs beyond the binding sites, such as overall amino acid compositions.

Based on these considerations, we defined three groups of features to identify novel DBPs:

(i) pDBS features: features derived from the prediction of DBS.
(ii) pOBS features: features derived from the prediction of binding site residues for other ligands.
(iii) AA-composition features: whole protein amino acid composition.

The sequence-based prediction of DBS assigns a numerical score to each residue in a protein. To define sequence features unbiased to the protein length, we calculated the following measures from the DBS prediction scores:

(i) The average score for the top five pDBSs.
(ii) The average score for the top 10 pDBSs.
(iii) The average score for the top 25 pDBSs.
(iv) Third quartile score for all pDBSs.

(These measures have significant redundant information, but a clear choice for the right feature was not obvious at the outset.)

A similar set of four pOBS features were derived for each of the four other ligands considered, i.e. ATP, RNA, other proteins and carbohydrates, thereby creating 16 additional pOBS features (40–43).

The AA-composition features are simply the relative number of each of the 20 amino acids in the protein chain and an additional feature of sequence length was added to this set.

Cross-validated training and prediction of DBPs from sequence. A multiple linear regression (MLR) model and a Random Forest (RF) model were independently constructed and a consensus was taken by averaging the prediction scores from the two models. (Taking the average of the two models instead of using only one of them was found to be more effective after several iterations of parameterization and model selections over small sample datasets, which tried to mimic the cross-validation strategy of the final approach on the entire data.) MLR was chosen as the simplest model for the relationship between training inputs and class labels (DBP versus non-DBP), whereas RFs were used to obtain the high fitting values. Other models such as non-linear support vector machines (SVMs) and neural networks were found to be computationally too expensive and to add little value to the predictions based on the combination of MLR and RF (data not shown).

To avoid overfitting and to select the best combination of features, we developed a 10-fold cross-validation scheme with feature-selection steps. By dividing the data into sets of 10% of SP_human proteins each, we merged nine sets at a time to derive a *training set* for selecting the best predictive features and model training. After feature-selection and training were completed, the predictions on the remaining 10% of the proteins (*test set*) were obtained using the trained model. The cumulative prediction performance over the entire set of proteins (collected from the test sets) was then evaluated. We used two methods for feature selection. First, we eliminated highly redundant features so that the filtered features included only those with a mutual Pearson correlation <0.95. Further feature selection was performed by a method known as 'recursive feature selection' as implemented in the Caret package of R (56), which simultaneously selects features and fits a trained model.

All cross-validation cycles of feature-selection and model fitting were performed by using various definitions of DBPs and averages from these models were used as the final prediction scores corresponding to each definition of DBP independently (see Supplementary Methods for DBP definitions).

Predicting DBP coding genes from global gene expression derived features

Compiling gene expression profiles. Gene expression omnibus (GEO) contains about 1.2 million gene expression profiles, accumulated over some 13,000 different 'platforms'. Of these, >2,000 platforms correspond to human samples (https://www.ncbi.nlm.nih.gov/geo/) (57). Despite continuing efforts (58–60), cross-platform data comparison poses serious problems, including batch effects, normalization and scaling. To avoid these issues, we decided to use

the transcriptome data from a single platform, GPL570 (Affymetrix platform with the largest number of available samples among the human platforms) for our analysis. Furthermore, to quantify the data uniformly across experiments without the assumption of normality, the absolute ELs in each *sample* were replaced with simple within-array ranks and only the samples with an identical number of probe-level EL values were retained, resulting in 72,488 *samples* or expression profiles. Hereafter, we use the phrase EL but it actually refers to the rank.

Compiling DBP annotations for genes in Affymetrix microarray chip. Probe-level annotations of the Affymetrix platform were converted to gene names by the dictionary provided in the GEO platform file. DBP coding genes were selected by virtue of their GO annotations, obtained from TargetMine data analysis platform (61,62). All levels of GO associations for biological process, molecular function and cellular component were utilized for this purpose, leaving the non-discriminating features to be eliminated to later feature selection and training models implicitly.

Computing gene expression level (EL) features. First of all, for a gene with multiple probes, the highest EL among these probes was selected. This treatment is consistent with the practice adopted in similar studies (46). Thus the EL features of each gene were calculated from these values. Two set of features were found to be useful: the average EL and the EL histograms as per pre-defined equal-probability bins on the entire data pooled together. Equal-probability bins come from $M \times N$ EL-values, where M is the number of genes and N is the number of EL-values for each gene. In this study, $M = 20{,}318$ and $N = 72{,}488$ as stated above. For the 20 equal bin values defined from a global pool of genes and samples, individual EL feature profiles for each gene were computed by counting the relative number of occurrences (out of M-values) in each of the 20 bins. These 20 values represent our EL feature set for each gene.

Computing co-expression level (CEL) features. To compute the co-expression features for each of the M genes, their EL values in N samples were compared with those for all the other genes. The resulting M-values (Pearson correlation coefficients over within array ranks) were summarized as co-expression histograms similar to the EL probability features described above. Again, the histogram bins were recomputed by considering all the $N \times (N-1)/2$ co-expression values from unique pairs of genes and the distributions of co-expression level (CELs) in these 20 bins were used as CEL features of that gene.

Computing network gene ontology composition (NGC) features. Network gene ontology composition (NGC) features were derived by computing a histogram of GO terms' occurrences; given a gene, GO terms for its T top co-expressed (positive correlations) and L least co-expressed (negative correlations) genes were pooled and counted. In this study, both T and L were set to 50, resulting in GO histograms based on the annotations of top 100 'co-expressed genes' for each query gene. Overall 138 GO terms were found to be present in at least 200 genes in the entire list (used for filtering) and hence the NGC features were composed of 138-dimensional integer valued vectors.

Predicting DBP genes from EL, CEL and NGC features. Similar to the cross-validation training strategy adopted for predicting DBPs from sequence features, we trained MLR, SVM and RF models by using EL, CEL and NGC features (together forming a 20 EL + 20 CEL + 138 NGC = 178-dimensional feature vector) as model inputs and the GO DBP gene assignment as targets. The 10-fold cross validation scheme described earlier was adopted to ensure no overfitting in the trained models.

RNA-binding protein (RBP) annotations. To assess if there was any potential contamination of RNA-binding protein (RBPs) in the predicted DBPs, we also included RBP annotations in our analysis. To achieve this, we simply retrieved all human proteins associated with the GO term *RNA-binding* GO:0003723 using TargetMine and assigned them as RBPs. We employed only the GO-based annotations for selecting the RBPs since they happened to be the most lenient of annotations and therefore, provided the largest repertoire of candidates for comparison.TargetMine contains robust annotation data for human, mouse and rat but this analysis can be easily extended to other species by making use of the GO annotations provided with UniProt data sets.

Training data for reproducibility and for other organisms

The feature sets used for training the models for *human*, *mouse* and *Arabidopsis thaliana* genomes are provided through our web site for this project (http://gigeasa.sciwhylab.org). Additional datasets and detailed protocols are also available on this URL. The binding site-based features employ the predictions using several of our previously published methods. We have also made available a document *how-to.pdf* to enable the users to understand the specific components of the predictive models developed for this work. To facilitate their computations for a new set of sequences, we have created an integrated web server to predict such features for hundreds of proteins in a single submission. Since the calculation of gene expression based features is computationally intensive, the standalone tools to compute them from a gene expression matrix are also provided on the above URL.

Integrated analysis of GE and sequence-derived predictions

To integrate the two approaches, the list of genes used in our GE-based method was compared to the UniProt human proteins, resulting in a total of 15,658 genes mapped to 15,710 proteins. The Uniprot-supplied gene/protein mapping and the Affymetrix annotations were used for this purpose. Prediction results for the genes mapped to multiple proteins were repeated for every protein entry to compare prediction performances. Sequence-based prediction scores were re-computed by excluding the proteins that did not map to any of the genes in the overlapped set.

Scoring model performances

Prediction performances of all the trained models were measured by calculating the area under the receiver operating characteristic (ROC) curves obtained by pooling together predictions from the test datasets of all the cross-validation iterations.

Motif enrichment analysis

For each candidate novel DBP, the top 50 positively co-expressed genes and an equal number of negatively co-expressed genes were collected as described above. For each gene in the two sets, DNA sequences within 1000 bases from their genomic location were collected and aligned using a well known software called 'MEME-CHIP' (63,64). From each run of MEME-CHIP, the significant motif as sorted by *E*-value was retained to examine their basic properties such as being a palindrome and the number of genes that contained that specific motif. The threshold of the motif search algorithm was fixed to be upto 10 bases with default parameters.

RESULTS AND DISCUSSION

The most significant results obtained in this study are summarized in three sections below: (i) results from the prediction of DBPs from sequence-derived features, (ii) prediction results from GE-derived features and (iii) integrated analysis of these results and estimates of cooperativity between sequence and GE-derived features. Detailed findings for each of the stated objectives are provided in Supplementary Tables ST1–10, Figures SF1–6 and Results SR1. Please note that most discussion in the following sections is related to the human genome, unless otherwise specified.

Prediction of DBPs from sequence features

A detailed statistical analysis of the 41 sequence features is shown in Supplementary Results SR1, which provides the rationale for their use in a predictive model.

Figure 2A shows the relative efficacy of different sequence-based feature sets in predicting DBPs (as defined by UniProt SeqFT). Clearly the amino-acid composition remains the best predictor, followed by features derived from predicted binding residues with their area under the curve (AUC) of ROC being 83% and 79%, respectively. An integrated model using all these features produced an AUC of ROC equal to 89% (see Supplementary Figure SF3c). Complete sets of prediction from this model, along with the DBP definitions for each protein, are provided in Supplementary Table ST9.

The highest scoring proteins can be classified into the following categories:

(i) True predictions as per the DBP bySeqFT annotations (DBP-by SeqFT).
(ii) Proteins that are annotated to be DBPs by one or more definitions, i.e. PDB, Pfam or GO annotations, but not bySeqFT.
(iii) Proteins with none of the known annotations as DBPs.

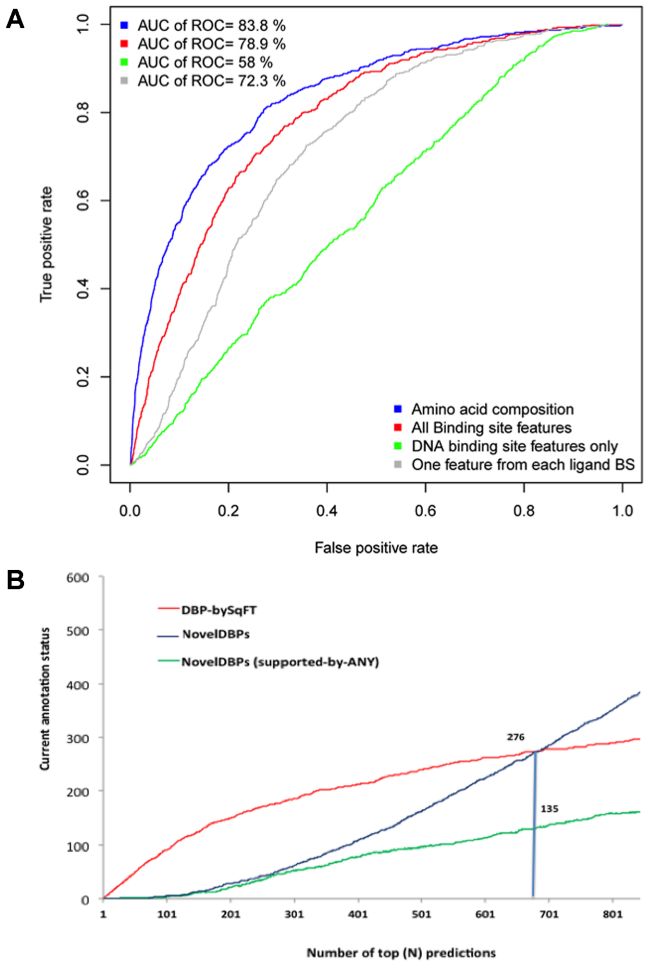

Figure 2. (**A**) Cross-validation prediction performance of linear regression models by different feature sets. Amino acid composition of the proteins was found to be the most effective predictor for DBP annotation. DNA-binding site residue prediction score summary, by itself is not very effective unless combined with binding residue prediction scores for other ligands such as RNA, proteins, ATP and carbohydrates (**B**) Estimate of the plausibility of apparent false positives to be novel discoveries as some of them are already annotated as DBPs by the critereon other than DBP bySeqFT. Of the top-scoring 687 predictions from SeqFT-trained models, 135 novel predictions are supported by at least one non-SeqFT annotation, while 50% of the remaining is true positives as per SqFT annotation. Additional data leading up to this Figure is shown in Supplementary Table ST9.

Figure 2B shows how the number of novel predictions, unannotated in UniProt SeqFT (groups 2 and 3 combined), changes with the number of known DBPs (group 1) or the number of unsupported predictions (group 3) at various thresholds applied to our final prediction scores. We chose a cut-off so that the total number of novel DBPs was equal to those already annotated as DBPs by UniProt SeqFT (corresponding to at least 50% confidence, if all novel predictions were false). At this cut-off, there were 276 DBP-bySeqFT proteins and an equal number of those not supported by SeqFT, of which 135 were completely novel (group 3), while 141 others are supported by one or more annotations from other sources (group 2) (Supplementary Table ST9).

We then examined the top predicted DBPs both in groups 2 and 3 for the likelihood of their engaging in DNA bind-

ing. First, we identified seven proteins (#1–7 in sheet #3 in Supplementary Table ST9), which had both Pfam and PDB-derived annotations as DBPs. They are in fact all well-known DNA binding proteins, and their missing SeqFT was rightly corrected by our analysis. The next 14 proteins (#8–21 in sheet #3 in Supplementary Table ST9) were annotated in PDB. This list contains several known DNA-binding proteins such as CCAAT/enhancer-binding proteins α and β (P49715 and P23771) and *Trans*-acting T-cell-specific TF GATA-3, which further supports the effectiveness of our method, because these proteins are generally missed by a standard UniProt search for the DNA-binding sequence features or even the Pfam domains associated with this function.

After inspecting these high priority hits individually, we surveyed other novel predictions and found that many of them are already known DBPs with varying degrees of confidence and knowledge, strongly endorsing the power of our predictive method and suggesting that many of our novel predictions may indeed have a DNA-binding function.

Prediction of DBPs from gene expression-derived features

To develop prediction models to identify DBP genes from GE profiles, we used three feature sets: (i) EL, (ii) CEL and (iii) NGC features (see 'Materials and Methods' section). Prior to their combined use in training the prediction models, we examined their statistical patterns to make a preliminary assessment of their ability to identify DBP genes.

Global expression levels of DBP genes. To estimate a general trend of ELs of DBP genes, we divided all observed EL values from all sample from GEO into 20 equal-frequency bins. The relative number of samples in which a gene was observed in each of these 20 bins was then counted (see 'Materials and Methods' section). A high frequency in lower (or higher) bins thus indicates that the gene is expressed poorly (or highly) in most samples. Figure 3A presents these data for all DBP and non-DBP genes, pooled together in the two groups. We observed that the distribution of DBPs is systematically skewed toward the higher ELs and peaks near the 15th of the 20 bins, implying that a larger number of DBP genes are expressed in ranges close to the 75th percentile of the global EL values than in lower EL ranges.

Figure 3B shows this distribution from a different perspective by simply taking an average of the ELs of DBP and non-DBP genes. The average ELs of DBP and non-DBP genes differ statistically significantly with a *P*-value in the order of $e\text{-}^{11}$. However, a very broad spread of data in the two categories and overlapping box plots in Figure 3B also indicate that this difference in the mean EL values is not very useful to distinguish individual DBPs from non-DBPs.

To examine whether a more detailed distribution of ELs will improve the predictive power of EL-derived features, we plotted a heatmap showing the distributions of ELs of each individual gene across the GE dataset (Figure 3C). The heatmap revealed that many (compared to non-DBPs) but not all DBPs are highly expressed across the samples.

To investigate this further, we sought to examine three different aspects: (i) what are the typical examples of highly expressed DBP genes, (ii) which DBP genes are poorly expressed on the average, defying the overall shift in DBP expression and (iii) what are DBP-like non-DBPs genes, which show an EL pattern that is similar to the majority of the DBP genes. Figure 4A–C explore the answers to these questions, respectively. In Figure 4A and B, we list the typical genes falling into these categories and investigate their general annotations, whereas in Figure 4C we perform a more comprehensive GO enrichment analysis. As shown in Figure 4A, many genes in the highly expressed category of DBPs are associated with housekeeping and ubiquitously expressed functions such as ribosomal assembly and ribonucleoproteins. Figure 4B shows that DBP genes that are most significantly different from general DBP genes and thus cannot be identified by a GE-based method are frequently associated with development and pluripotency. Figure 4C supports the observations from Figure 4A and B and identifies additional GO terms that can characterize genes (either DBPs or non-DBPs) identifiable by global GE analysis. In summary, we conclude that even though global ELs are not sufficient to categorically distinguish between DBPs and non-DBPs, there exists a significant bias that may potentially be exploited in combination with additional knowledge about these genes.

Global co-expression level (CEL) and network GO composition (NGC). To further identify the differential GE features between DBPs and non-DBPs, we first compared the overall CELs of each DBP and non-DBP gene with all the other genes. Basic comparisons of these co-expression profiles (CEL features) are shown in Supplementary Figures SF4 and 5. Even though DBP genes were found to have overall higher CELs than the non-DBP genes, differences in the CEL features were relatively small. However, a stronger signal for DBP genes was found when we identified the top and bottom 50 genes co-expressed with each gene and computed the enrichment of their GO terms (see 'Materials and Methods' section). We found that the GO composition of genes co-expressed with DBPs (network genes) were significantly different from the GO composition of the non-DBP network genes. Figure 5 shows a summary of the comparison of GO terms, which were found to be most significantly enriched in DBP networks compared to the non-DBP networks. Many GO terms such as 'nucleus (GO: 0005634)' and 'DNA-templated transcription (GO:0006351)' and terms intuitively related to DNA-binding were enriched in the DBP gene co-expression networks. On the other hand, certain GO terms such as 'integral component of membrane (GO:0016021)' were depleted in the DBP gene networks compared to the non-DBP gene networks. In other words, a DBP gene is more likely to co-express with other genes with similar GO annotations. This phenomenon leads to the enrichment of these GO terms in the co-expression network. Genes that are less likely to be co-expressed with the DBP genes lead to a depletion of certain other GO terms. This result is not surprising as DBP genes often associate with co-factors and specifically and contextually co-localize with other genes to enable their function. This finding is useful in computing a GE-derived NGC features for predicting DBP genes.

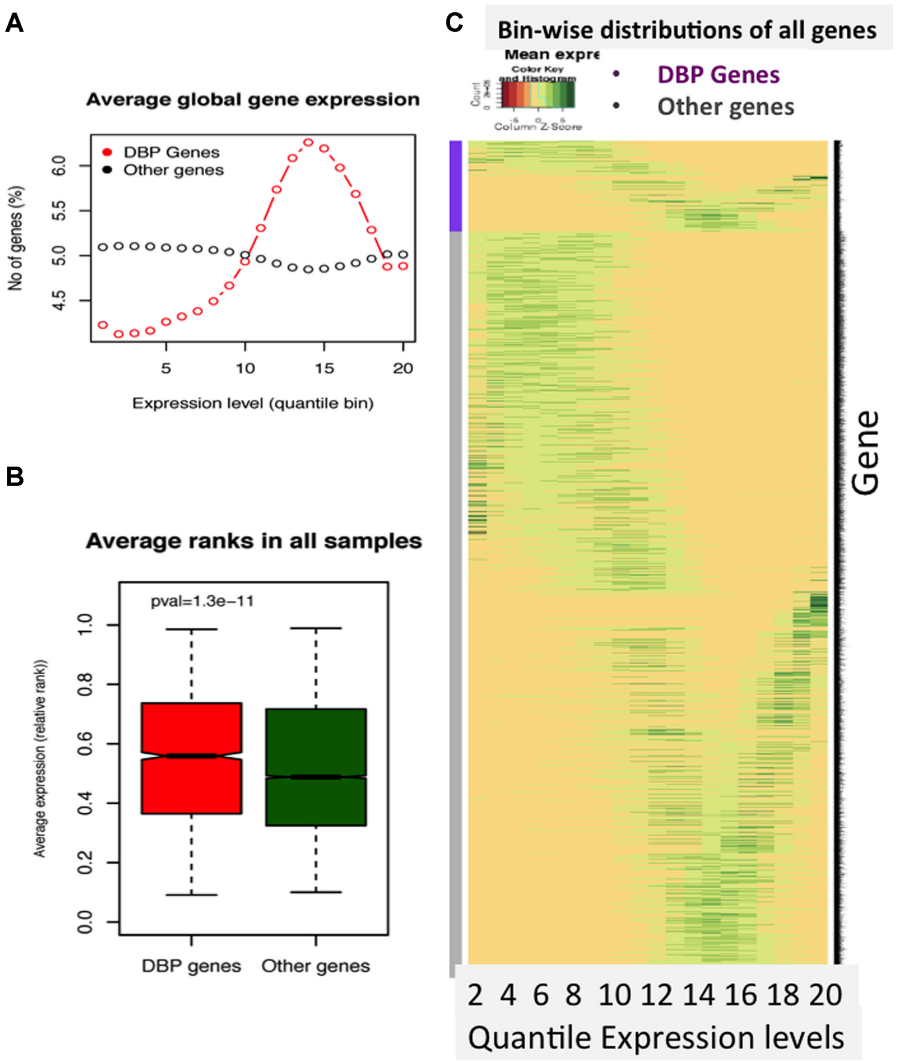

Figure 3. Distribution of DBP and non-DBP genes among the 20 bins of average global ELs. (**A**) Even though the DBP genes are distributed in all these bins, their distributions are skewed toward higher ELs. (**B**) The same data as in (A) seen as gene-wise average EL confirms these observations. (**C**) A detailed frequency histogram of expression values of *each* gene in the 20 global quantile bins. In the heatmap, each row represents a gene and columns are 20 bins of their EL (increasing order). The colors annotate the relative number of times that a gene is expressed in the EL range described by the column. The heatmap corroborates the conclusions in (A) and (B).

Assessing cumulative power of gene-expression derived features. As we noted above, the three feature sets derived from the GE profiles of DBPs carried useful information about their annotations. Their cumulative ability to predict DBP genes was estimated by training machine learning models using all of the features as inputs. Figure 6A shows the results from three independently trained machine learning models (SVM, RF and Logistic Regression) using a 5-fold cross-validation scheme (see 'Materials and Methods' section). The prediction scores from these three methods were combined to develop an all-model consensus by taking a simple average. As seen in Figure 6A, the all-model consensus worked slightly better than individual models, which showed very similar levels of accuracy. Specifically, the AUC of ROC for RF, SVM_radial, Logistic Regression and all-model consensus was 63.90, 62.74, 63.92 and 66.0%, respectively. In contrast, the AUC of ROC for the sequence-based prediction results mapped to a common dataset was 85.46%. Thus, the sequence-based prediction of DBPs comprehensively outperformed the GE-based prediction.

Integrating sequence and gene expression based predictions

We then examined whether the sequence-based prediction of DBPs could be improved by combining it with the GE-based prediction scores. We adopted two approaches, as explained below, to combine the sequence and GE-based predictions and their results are shown in Figure 6B, together with the prediction results before combining. Note that the GE data are not available for all the proteins used in the sequence-based predictions and vice versa, so these prediction performances were re-computed for the common dataset.

A simple consensus obtained by combining the sequence and GE-based prediction scores did not improve the high performance of the former model. However, a manual in-

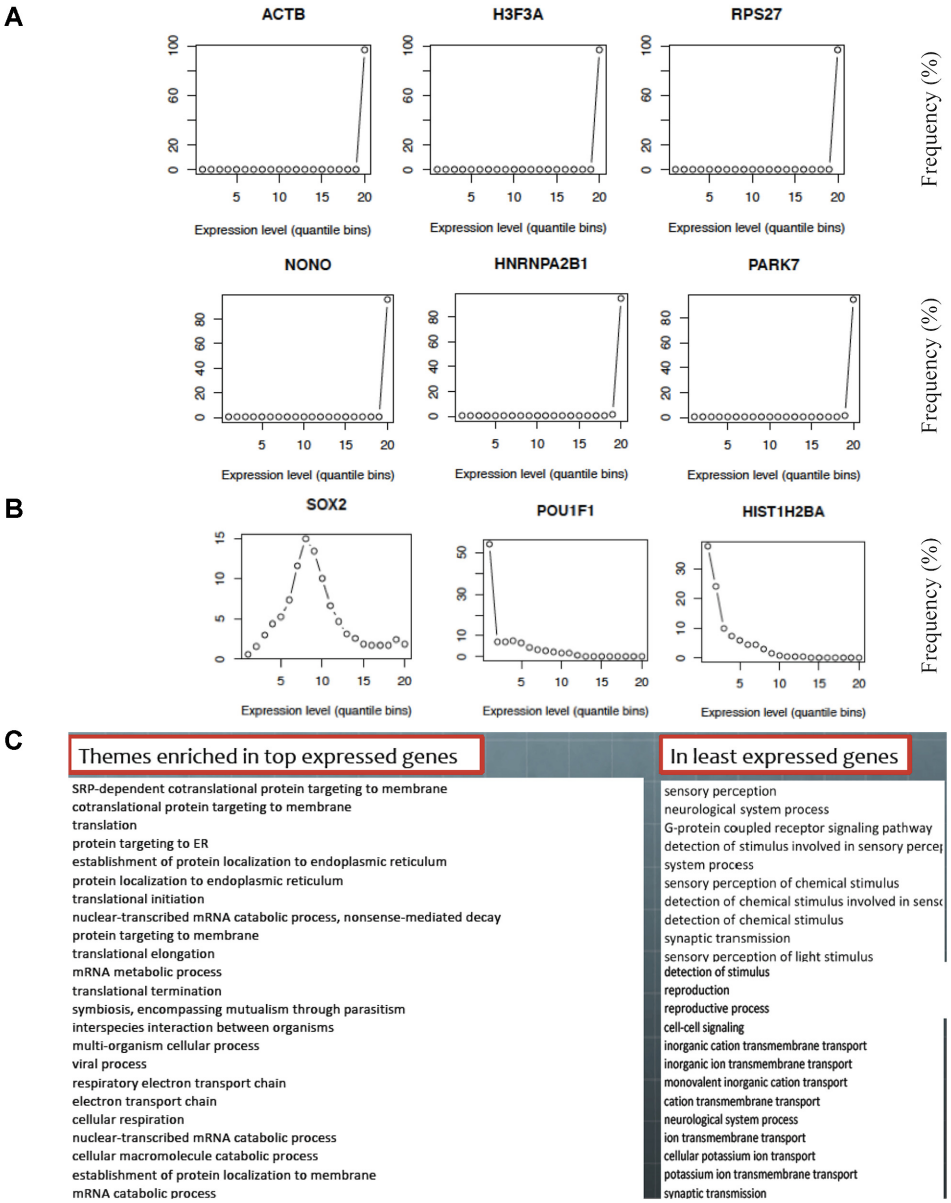

Figure 4. Typical global gene EL distributions of selected DBP genes and the co-occurring biological themes: (**A**) Many DBP genes, specially associated to ribosomal assembly are ubiquitously highly expressed as their ELs (*X*-axis) are almost always (*Y*-axis) in the top bins. (**B**) Some well-known development-associated and other DBPs are expressed less frequently and at a lower EL. (**C**) Specific GO terms are enriched in top and bottom expressed DBP genes, providing a clue useful for their predictions.

spection of the uniformly rescaled prediction scores from the sequence and GE models for DBP genes revealed that GE scores were particularly high for DBPs when their sequence-based scores were below a certain level ($Z < 3.5$). On the other hand, GE scores were low for DBPs with high sequence-predicted scores and therefore, added little substance to the sequence-based prediction. Based on this observation, we have proposed a conditional consensus, wherein the GE-predicted scores were combined with the sequence-based prediction only if the sequence scores were below the specific cut-off. This heuristic-based conditional consensus bested the sequence-only and simple-consensus based prediction scores, resulting in the final AUC of ROC of 86.5% for the common dataset.

Biologically, it suggests that high sequence scoring DBPs would likely perform that function irrespective of gene expression environment, wheres there are other DBPs, which carry a weak sequence-based DNA-binding signal. This group may consist of proteins, which do not bind to DNA on their own and need assistance from co-factors or those that can bind to DNA only when over-expressed. (Alternatively, they may be co-factors themselves, which contain a weakly interacting DBS, as predicted by sequence information.) We, therefore, investigated the correlations between functional categories of DBPs and the extent to which their GE features can improve the sequence-based prediction performance. GO terms with sufficiently large populations of DBPs and non-DBPs were collected and individual

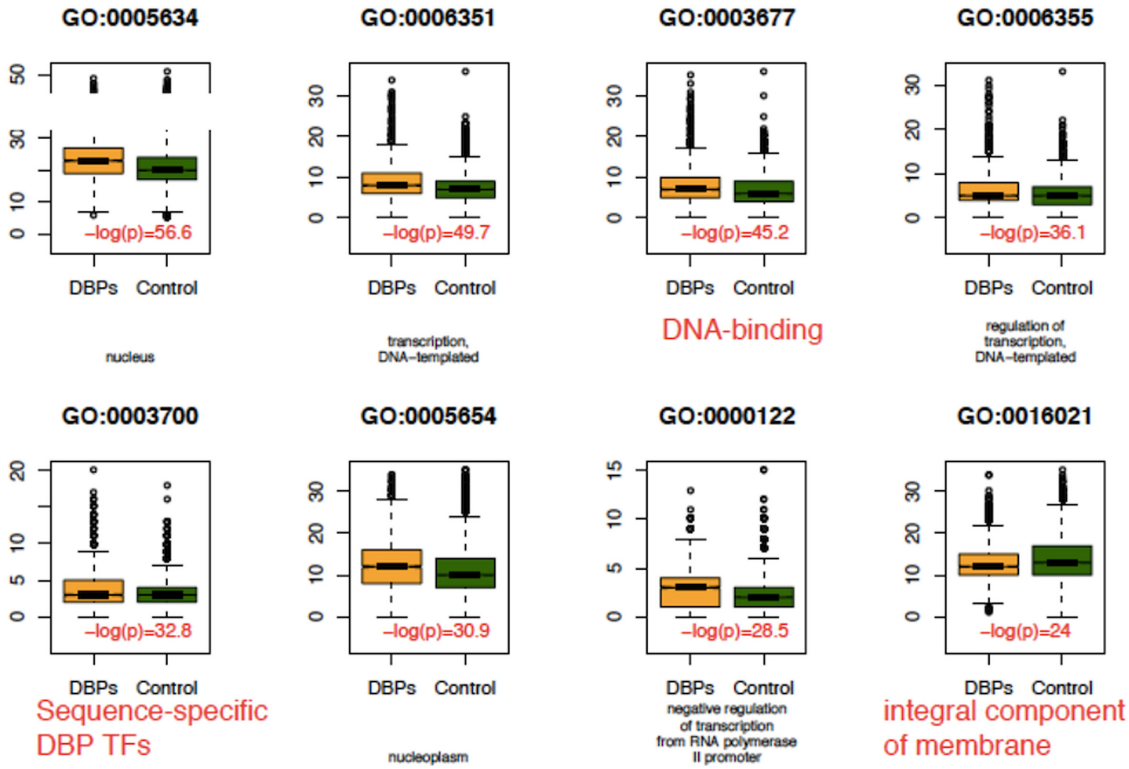

Figure 5. Most significantly enriched/depleted GO terms in the top and last N genes co-expressed with DBP genes. $N = 50$ for the current plots. *P*-values in the negative $\log_{10}$ scale are written in the insets.

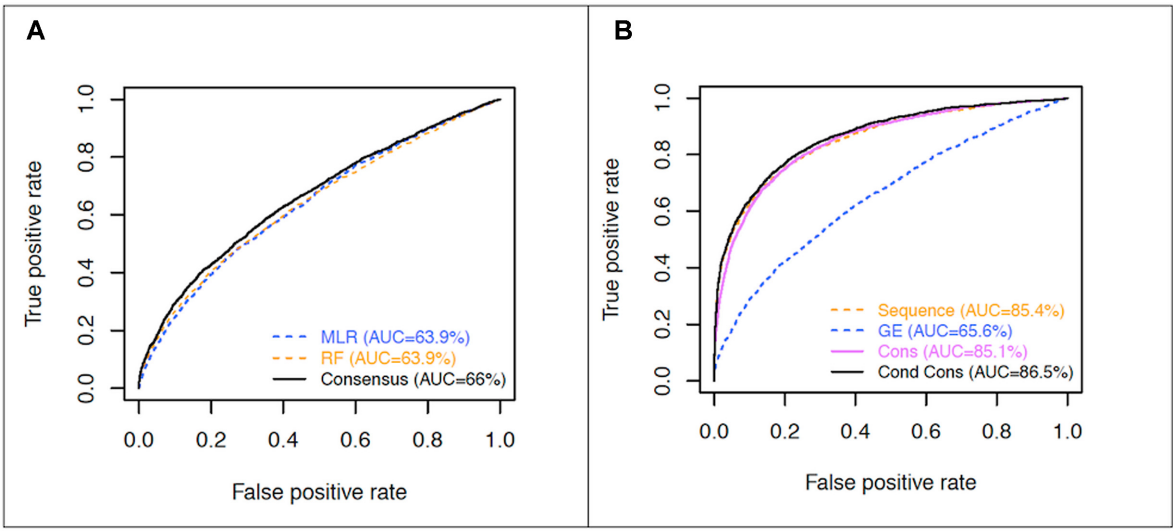

Figure 6. Gene expression features and DBP predictions: (**A**) DBP prediction performance of gene expression-based models using simple and complex computational models and their consensus (**B**) Comparison of sequence-based prediction models with GE-based models on the overlapping gene list and the ability of the consensus to improve prediction performance over the two methods. Performance could only be improved by a conditional consensus where GE-based scores are added to only those genes in which sequence-based DBP prediction score was low, suggesting a compensatory or cooperative mechanism of DNA-binding activity.

groups associated with each of them were analyzed. Gains in the prediction performance upon taking a simple consensus for the genes corresponding to each GO term were computed. A histogram of these performance gains across GO terms is shown in Figure 7A.

Figure 7A showed that for most GO terms, the performance of sequence-based models is not affected much when a consensus with GE-derived features is included. However, a few GO terms showed a large variation in performance, by either working in tandem with the sequence-based prediction by increasing the prediction performance or by impacting them negatively through the generation of potentially contradictory prediction scores. The GO terms in these two extreme groups are plotted in Figure 7B. As shown, the GE-based prediction was found to be most useful for genes associated with *metabolic process* (GO:0008152), as the AUC of the ROC for 361 genes in this group was merely 62.7% from the sequence-based models. The GE-based prediction for this group was 87.2% and the consensus performance was 78.0%, clearly indicating their superiority in identifying these DBP genes over the sequence-based models.

There are a few other GO terms such as those associated with GTPase and ATPase activity, which do not show large differences between sequence and GE-level performances but nonetheless potentially contribute to the DNA-binding functions of these genes. On the other hand, GO terms such as Golgi membrane, ER membrane and even sequence-specific DNA-binding TFs are apparently best predicted by sequence-information and GE-features negatively affect their predictability.

The two observations presented here, i.e. the DBPs with a poor sequence signal are better predicted by GE features and the differential performance gains of genes with certain GO terms, suggests an interplay between sequence-encoded functions and the cellular contexts defined by GE patterns. Our results suggested that the GE dynamics can compensate for low DNA-binding signal at the sequence level. Presumably, the proteins coded by these genes have relatively fewer DNA-binding residues that are probably insufficient to recognize DNA at the same level as other DBP genes. Greater EL values, the presence of co-expressed genes aiding their DNA-binding activity and the functional dynamics of co-expressed genes, which are the basis of GE-based predictions in this work, may likely facilitate the DNA-binding ability of these genes.

Alternative prediction models and RNA-binding protein contamination

To examine if the proposed model performance could also be obtained from alternative computational approaches, we investigated two additional approaches to evaluate how the current model performs in comparison. Furthermore, we also evaluated the extent of 'contamination' of the signals from RBPs in our DBP prediction models, which is frequently a confounding factor in such models (28).

Post-training/pre-training combination of sequence and GE features. In this study, we have trained the prediction models with GE-based and sequence-based features independently and integrated them post-training. To test if a pre-training integration of features and training models with all the features together would perform better, we also trained models by taking all the features together. Figure 8A shows the results of various models trained by an integrated feature set made up of all the 175 computed features (using GO-based DBP annotation as the target class) compared with the post-training integration models described above. We observed that the RF is the best computational model for training all the features together. Figure 8B shows that the strategy presented above, i.e. training the sequence and GE based models separately and integrating them post-training, outperforms all the other models that employ feature-level pre-training integration. We also developed several training models with different combinations of feature sets and parameters in feature selection and found that the difference caused by tweaking the feature size is about 1–2% points in terms of AUC (see Supplementary Figure SF5).

DBP prediction using sequence alignments. The results presented in this study lend themselves to the question if the elaborate models presented here could be replaced by a simpler method to annotate a new DBP if it shows a sequence similarity to other known DBPs. The caveat remains, however, that the sequence similarity between proteins may occur both within the DBP and non-DBP regions and for many DBPs the actual DBS may not be known. This consideration puts our proposed model at an advantage because it is driven by the binding site prediction at the sequence level.

To evaluate this issue in detail, we performed a 10-fold cross validation on alignment-based DBP prediction, i.e. divide the data into 10 parts, pool 9 of them, create a database of DBPs from these pooled data and find the best matching DBP sequence (based on the *e*-value) for the tenth fold test data and assign the DBP prediction score from this e-value. These data pooling from nine of the parts and predicting DBP for the 10th part is repeated for all the 10 parts, leading to the predicted DBP scores for each protein in our dataset. Figure 8C shows the prediction results of such an approach. The integrated models clearly outperformed the alignment-based approach by 1.5 percentage points. This difference may appear trivial at first, but Figure 8C shows that our integrated models have a substantial advantage in the low false positive range of predictions. For example, with a 5% false positive rate, our integrated model-based predictions achieve a true positive score about twice as better as the alignment-based predictions, thereby, providing a 100% improvement. In terms of the actual number of predictions made, this is substantial, as the false positive rates are relative to the large negative data and 10% of the false positive rate would amount to 2000 proteins! Moreover, the low false positive range is in fact the actual prediction range of interest, as there are likely to be very few true cases per false prediction if the false predictions are too high.

In the alignment-based predictions, the first point on the ROC curve is observed at a significant number of false positive. This observation is due to the fact that the first true positive hit by the method of sequence alignment is also accompanied by a large number of non-DBPs aligning to the DBPs in the reference data, presumably in the non-binding regions and the non-DB domains, which are often

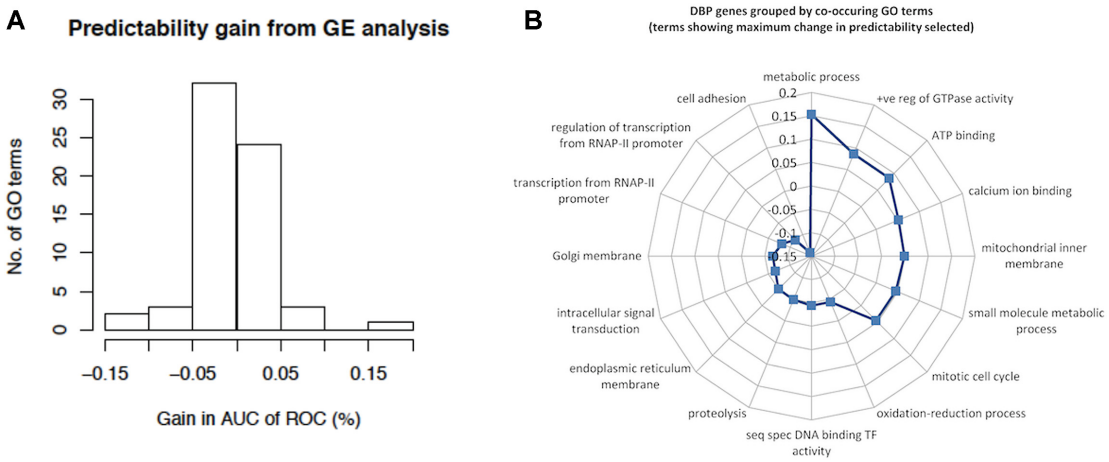

Figure 7. Improvement in sequence-based DBP prediction by adding GE-based prediction scores in DBP subgroups with co-occurring GO terms. (**A**) Number of GO terms, showing a specific range of improvement (gain in AUC) in prediction (**B**) most significantly impacted group of GO terms and corresponding improvements in prediction performance. Gene-expression related features appear to be crucial for DBP function such as metabolic process, as they have low sequence-based signal for DBP function.

co-located in the DBPs. We examined some of these false positive cases and did find a number of such obvious instances. For example, several TLRs, which do not have a DNA-binding function, align well to other TLRs in their transmembrane domains, thereby showing up as false positive hits in the alignment based searches. It is therefore evident that *a priori* annotations of the DNA-binding domains are required to construct a robust reference dataset for DBP annotations based on alignments. Such detailed annotations are available only for a small number of DBPs and hence the proposed method, with no requirement of such reference datasets, is at a clear advantage.

The current models using sequence and GE in the structure employed here appear to outperform other models (see also section on 'Role of redundancy in performance'). However, this conclusion could also be so because we have optimized only one such approach and we do not rule out that an improved approach to combining sequence alignments, GE and predicted binding site based method may perform better.

Contamination of RNA-binding proteins in predicted DBPs. A typical concern that has been expressed in recent studies is that DBP prediction methods often cross-identify RBPs as DBPs (28,65–66). In our approach, we have incorporated both the DNA- and RNA-binding site scores with the hope that the latter could serve as a background score and thereby improve specificity of predictions (40–43,49). To evaluate the degree of RBP annotations at various thresholds of DBP prediction score, we computed RBP prediction rates for every fraction of DBPs predicted at different cut-offs. Figure 8D shows the AUC of ROC for these erroneous predictions and Figure 8E shows the fraction of RBPs at various levels of true DBP predictions. Again, we observed that the model performs very well in discriminating DBPs and RBPs in the top prediction score range, as a very small proportion of RBPs were predicted to be DBPs in this range. For example, 40% DBPs could be detected at a cut-off, which also picked up about 10% of RBPs in the final predictions. At a more relaxed threshold, when many more DBPs are selected, the number of predicted RBPs as DBPs also increased, but since a high false positive range is less desirable as explained above, we concluded that our approach is successful in distinguishing many RBPs from DBPs.

Motif enrichment analysis

Our work is focused on identifying novel DBPs irrespective of their detailed functions such as TFs, histones, DNA sensors or enzymes. While the functional analysis of each of these categories is beyond the scope of this work, we decided to analyze the enriched motifs in co-expressed genes, assuming that the proposed DBP, if it were a TF, could potentially target these genes. We primarily focussed on genes that at the time of this study were assigned only an open reading fam (ORF) and not proper gene names and had no or little functional annotations in UniProt (see Supplementary Table ST10). For each of the top 100 proposed novel DBPs, the top 50 positively co-expressed genes (with highest Pearson correlation) were gathered. Another group of top 50 negatively co-expressed (with highest negative correlation) genes were also selected. For each of the co-expressed genes in the each of the two datasets, DNA sequences within 1000 bases upstream of their transcription start sites were retrieved and aligned using 'MEME-CHIP' (63,64). Supplementary Table ST10(b) contains the summary of the top motif detected by MEME-CHIP for each candidate DBP. As the table shows, enriched motifs that are present in as many as all the 50 co-expressed genes are observed for some candidate DBPs. A detailed investigation and characterization of individual motifs is beyond the scope of this work, nonetheless, our observations have provided useful insights into the potential functions of these ORFs and are well primed to facilitate their further characterization. Among specific examples, we observed two candidate ORFs namely C6ORF23 (CCTGGG in 44 of the top 50 positively correlated genes) and C6ORF15 (CCAGCCTGG in 10 of the

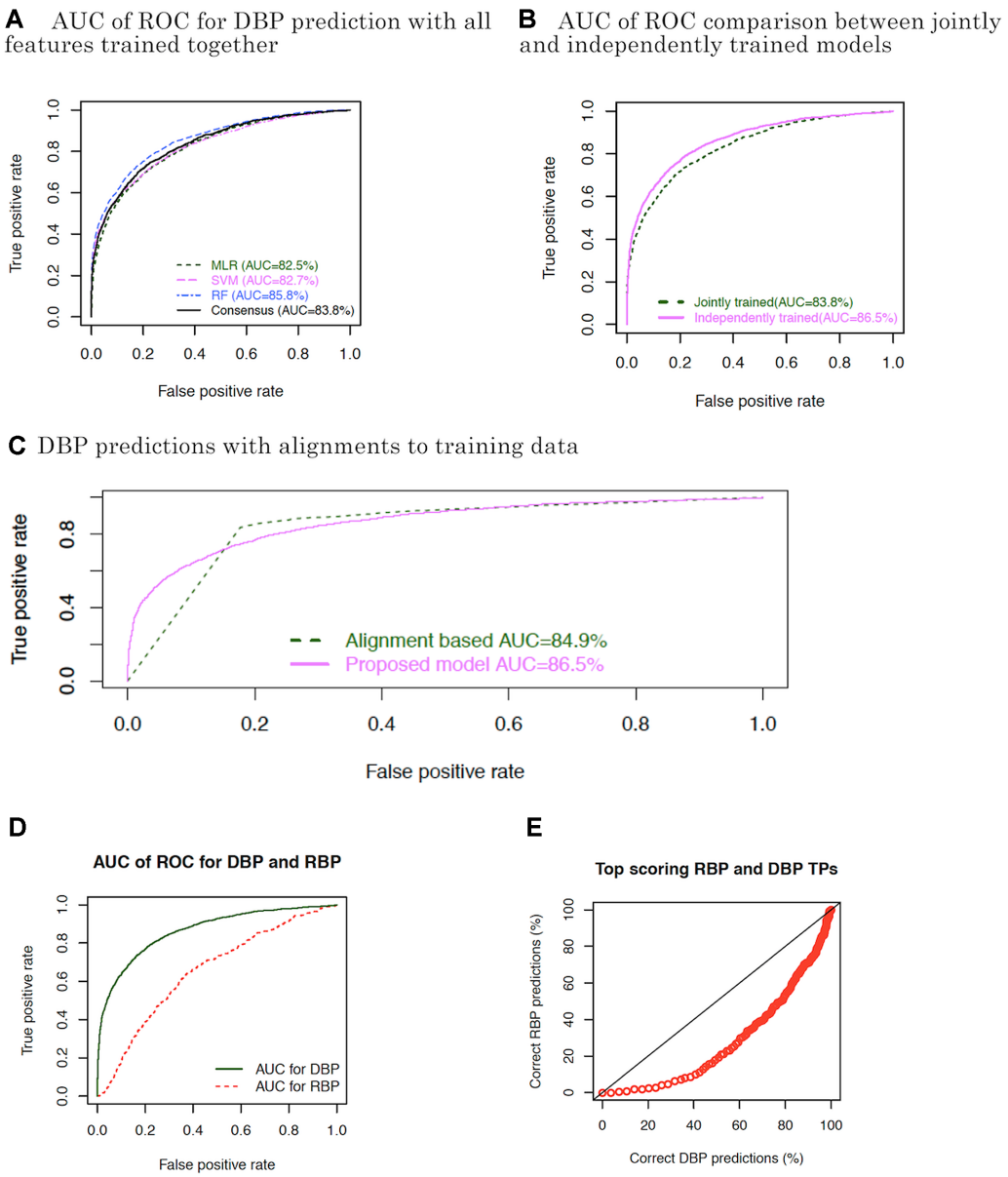

Figure 8. Evaluating the DBP prediction performance using various predictive approaches. (**A**) Using jointly trained features versus proposed independent training (**B**) sequence-alignment based annotation. (**C**) DBP prediction using sequence-based alignments is less effective than integrated models. (**D**) RBP prediction as a fraction of predicted DBPs at different cut-offs. (**E**) fraction of RBPs at various levels of true DBP predictions.

top 50 positively correlated genes), whose consensus motifs resemble those of known TFs in terms of palindromic sequences separated by a linker. In the future studies, we propose to undertake a more detailed characterization of these candidate DBPs.

To examine if C6ORF23 and C6ORF15 could be assigned putative TF annotations based on homology-based functional transfer, we performed a sequence search of Uniprot, SwissProt database (Human) for C6ORF23 and C6ORF15 homologs using the SSEARCH program (67). SSEARCH did not pick up any known TFs as close homologs of either C6ORF23 or C6ORF15 at an *E*-value threshold of <0.05. However, SSEARCH did identify SF3A2_HUMAN (Splicing factor 3A subunit 2), an RBP, as a homolog of C6ORF15 with marginal statistical significance ($E = 0.087$). Taken together, our results suggest that it would have been difficult to assign a definitive TF annotation to C6ORF23 or to C6ORF15 based on homology-based approaches alone, which became possible using our proposed approach.

Applicability of results to other genomes

Our study is primarily based on the annotations of human genomes and the training exclusively of the corresponding datasets. To evaluate whether the current models were also applicable to other genomes, we performed the sequence, GE and integrated predictions for mouse and *Arabidopsis thaliana*, the organisms with the next highest number

of reported experiments in GEO. Keyword annotations in UniProt have undergone significant updates since our first analysis on the human proteome, started much earlier in this work. Thus, we limited our analysis of the two genomes to the other four annotations, namely, SeqFT in Uniprot, GO, Pfam and PDB. Figure 9 summarizes the results from these predictions and the final predicted scores are available through the *gigeasa* project web site (gigeasa.sciwhylab.org). Key findings from such models were as follows:

(i) Barring PDB-based annotations, the prediction performances across human, mouse and *Arabidopsis thaliana* follow similar trends. PDB-based annotation was particularly poor in the case of the *Arabidopsis thaliana*, because very few (only 26 in our dataset) DBPs could be annotated by this method in this genome. Apparently most DBPs in *Arabidopsis thaliana* have not been characterized structurally and share poor sequence similarities with the existing protein–DNA complexes in the PDB.
(ii) Gene expression alone achieved a prediction performance of 66.7 and 68.1% for *Arabidopsis thaliana* and mouse, respectively, which are comparable to the results obtained for the human genome.
(iii) While both mouse and *Arabidopsis thaliana* show a consistent level of improvement on applying conditional consensus as against sequence-only or the simple consensus approach, the gains are relatively small (0.5% points) in the latter compared to human and mouse, both of which showed an improvement of >1% point. This variation may be due to the conditions imposed (adding the GE at a fixed cutoff) were optimized for human and adopted without adjustments to the other two genomes. This could also be due to better GO annotations available for human and mouse or additional factors that need to be further investigated.

Taken together, the models based on conditional consensus appear to work the best in all the three cases considered here and could even be improved in the future studies.

Role of redundancy in performance

The results presented above were obtained by training data taken from one entire genome at a time, wihout filtering for similarity thresholds. This obviously leads to redundancy in the datasets. Consequently, the presence of similar proteins in the training and test datasets may lead to overestimates of prediction performance levels. It may be recalled that even though it is customary to train prediction models after removing redundancy, there are instances where entire available datasets have been used. Despite our own protocol earlier to the contrary, we believe that the identification of novel DBPs need not be performed with a fully non-redundant dataset and maximum amount of information, even if somewhat redundant should be allowed in the trained models. However to gain a more realistic estimate of the prediction performance, we retrained our models of the pooled data and the human genome at various degrees of non-redundancy. We ran *blastclust* and selected a representative from each cluster computed at 25, 50 and 90% sequence similarity and retrained our models. Figure 9B and C show the results from these predictions. Two points are immediately noteworthy in these plots. First, the removal of redundancy from 100 to 25% sequence identity thresholds leads to a fall in performance by about 5% points, but even at 25% sequence identity threshold, our models retained their strong performances thereby demonstrating their predictive prowess. Obviously, this is a huge advantage over sequence alignment based annotations discussed above. Second, we observed that even at the redundancy removal at 25%, the pooled data performance is higher than the human-only data. This may be because the pooled data are more diverse and have more predictive power but may also indicate that some degree of redundancy remains despite the strict cut-off. We leave this issue to be resolved in the future work, but it is clear that the performance levels in each approach are strong enough to support the conclusions drawn in this study.

Related studies in the public domain

There are nearly 30 published studies, which report a method to predict DNA-binding proteins from sequence or predicted structure, whereas many more predict DNA-binding site residues. We have compiled a list of all such studies in Supplementary Table ST12a–c. To the best of our knowledge, there are no studies that have attempted to integrate gene expression profiles and sequence information for predicting DBPs. Moreover, the integration of DNA-binding site residues and DNA-binding proteins has not been attempted in a comprehensive manner so far. The only study where such an effort was attempted performed only a count of the predicted DNA-binding residues. Furthermore, in most of the studies, DBP datasets were compiled from protein–DNA complexes in the PDB and models were evaluated on various degrees of data redundancy and with different performance scores. There is only one study (68) that was based on first predicting a protein–DNA complex, in which an entire human proteome was the subject of prediction. In general, prediction performances have approached close to 0.80–0.90 in terms of AUC and 80% accuracy (the number of correct predictions in a mix of DBPs and negative controls). Due to high variance in the background datasets and DBP annotations, it is unclear as to which of these methods would perform well on the datasets that we have considered in this study. It is possible that some of the binding site prediction methods other than those used in this work perform better in predicting DBPs. However, the primary goal of this work was to establish a rational approach to combining binding sites and binding proteins and to evaluate if the gene expression-based features would add value to such predictions. Therefore, an apparent cooperative mechanism toward DBP function, as suggested by our work, generates confidence in the presented findings and opens up new avenues to investigate context-specific functional annotations of DBPs.

Even though the benchmarking of the effectiveness of each binding site prediction method neither feasible nor desirable as above, it is worth evaluating whether the novel feature sets of DNA-binding site statistics and those from other ligand binding sites play any role in prediction per-

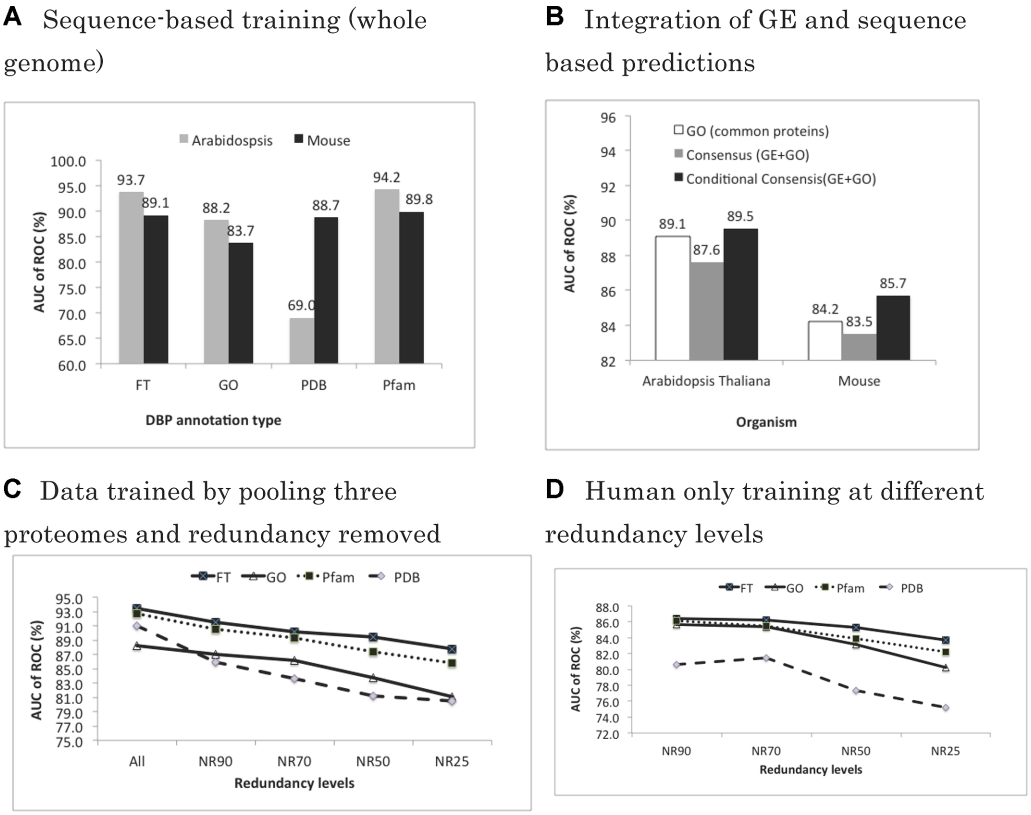

Figure 9. Various prediction performance scores for sequence based models and their counterparts with gene expression integrated scores in proteoms of different organisms. (**A**) Sequence-based training (whole genome) (**B**) Integration of GE and sequence based predictions (**C**) Data trained by pooling three proteomes and redundancy removed at 25, 50, 70 and 90% sequence identity thresholds. (**D**) Human only training at different redundancy levels.

formance. A clear answer to this can be better obtained by performing controlled experiments on training and benchmarking instead of published web servers where many of the proteins being predicted may have been variously included in training. Thus, we re-trained our prediction models by dividing our feature sets into three subgroups viz: (i) features derived from predicted DNA binding sites (ii) derived from binding sites from other ligands and (iii) amino acid composition of the whole protein. We computed the prediction performances of trained models by excluding (ii) and (i) to estimate their effectiveness in improving prediction results. Table 1 shows the results from such experiments under various redundancy conditions and DBP definition. We observed that the amino acid composition remains a good predictor of DBPs as reported earlier. DNA-binding site based features alone are weak predictors but when combined with the background binding site features of other ligands and amino acid composition, they outperform the other models lacking such features by a statistically significant margin (based on the *P*-value) over the distributions of prediction performances.

CONCLUSION

In this work, we have established a comprehensive system to predict DBPs from the amino acid composition and the predicted binding-site residues. Simultaneously, we also developed DBP prediction models from the global GE data. Even though the GE-based predictions were less accurate, we argue that the GE data provide greater value over their sequence counterpart, as they can introduce context-specificity into the computational models. Combining GE-derived signatures with sequence-encoded annotations will open up exciting possibilities of context-specific functional annotations. We observed that the many proteins associated with certain GO terms e.g. genes associated with *metabolic process* (GO:0008152) were better predicted as DBP or otherwise by combining the prediction scores with GE than the sequence models alone. Low sequence-based signals for such proteins could be successfully boosted by the GE-based prediction models, allowing a more accurate prediction without *a priori* knowledge of their GO term associations. We hypothesize that the GE patterns and network properties of DBP genes can function as an additional regulatory mechanism by not only defining a given biological function but also assigning a DBP function in a conditional manner. Many instances of multiple and independent biological functions of a single gene have emerged (69,70) and it is impossible to evaluate experimentally the conditional functional diversity for each DBP gene. While we do not address the issue of context-specific functional annotations in this work, we believe that a detailed analysis of GE patterns of DBP genes (as presented here) will help establish general principles of obtaining GE-derived annotations, which can be extended to context-specific annotations in the future.

Table 1. Sequence-based prediction performance of DBP prediction under various redundancy conditions and DBP definition

	DBS only	AA composition only	AA + DBS	All features
NR90-FT	63.50	86.00	86.20	86.40
NR70-FT	63.00	85.60	84.90	86.20
NR50-FT	61.30	84.40	84.90	85.30
NR25-FT	63.00	82.80	83.30	83.60
NR90-GO	64.60	85.60	85.50	85.60
NR70-GO	65.00	85.40	85.60	85.40
NR50-GO	64.70	82.60	82.90	83.10
NR25-GO	62.30	79.90	80.00	80.20
NR90-Pfam	61.90	85.40	85.30	86.10
NR70-Pfam	61.10	85.00	85.40	85.50
NR50-Pfam	61.60	83.20	83.80	83.90
NR25-Pfam	62.70	81.40	81.80	82.20
NR90-PDB	56.60	79.70	79.80	80.60
NR70-PDB	53.40	80.30	79.50	81.50
NR50-PDB	55.60	77.70	76.70	77.30
NR25-PDB	55.60	75.90	75.90	75.20
Mean	60.99	82.56	82.59	83.01
P-values (*t*-test)AA + DBS versus all features				0.0058

Feature set based on amino acid composition is the best of the three in most cases. However, adding DBS features to the model improves its performance in almost all the prediction models with a statistical significance in the improvement being observed by a *P*-value of 0.0058 by *t*-test. Abbreviations: FT: DBP definition taken from Uniprot Sequence features, NRxx: Data are non-redundant at xx% sequence identity threshold, GO: DBP definition taken from GO, DBS: DNA-binding site predictions, AA composition: amino acid composition of the full length protein.

Our results provide a framework to generate condition-specific DBP annotations by creating condition-specific GE profiles and establishing their roles in DBP functions. We believe that the functions of other groups of genes can also be readily fine-tuned by their gene expression patterns. However, detecting such an influence (or a lack of it) in all categories of proteins is much more challenging and has not been attempted in this work.

In this work, we have focussed on DBPs in general. However, a more specific functional prediction will be helpful for these DBPs and thus, our future works will be aimed at this broader objective. In the current work, we have examined the potential of top predicted candidates to be TFs. Of the top 100 scoring candidates that were examined, two, C6ORF23 and C6ORF15 were hypothesized to be novel TFs that may recognize predicted DNA motifs that closely resemble those of known TFs. A more elaborate analysis of the functional annotations described in this study is underway.

Finally, we have made our prediction results available on the web together with tools and data sets for computing features used in this work. This integration will enable users to examine the predicted DBPs in the context of their own gene set analysis with the associated biological information. To this effect, we plan to incorporate our predicted DBPs into the TargetMine data analysis platform (61,62).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Authors' contributions: This study was designed and implemented by S.A., with help from AA in performing benchmarks and data processing. P.P. contributed to the MEME-CHIP analysis. All other authors, led by K.M., contributed through critical discussions and helped in improving the manuscript.

FUNDING

Japan Society for the Promotion of Science Scientific Research Grants-in-Aid [15K40019 to S.A., 25430186, 25293079 and 17K07268 to K.M., in part]; Ministry of Health, Labor and Welfare (MHLW); Research on Development of New Drugs, the Japan Agency for Medical Research and Development (AMED) ('The adjuvant database project') [16ak0101010h0005 to K.M.); Department of Science and Technology, India DST-PURSE Grant; University for Potential of Excellence (UPoE-II), JNU, India [#270 to S.A.].

Conflict of interest statement. None declared.

REFERENCES

1. Man,S.M., Zhu,Q., Zhu,L., Liu,Z., Karki,R., Malik,A., Sharma,D., Li,L., Malireddi,R.K., Gurung,P. *et al.* (2015) Critical role for the DNA Sensor AIM2 in stem cell proliferation and cancer. *Cell*, **162**, 45–58.
2. Li,Y., Varala,K. and Coruzzi,G.M. (2015) From milliseconds to lifetimes: tracking the dynamic behavior of transcription factors in gene networks. *Trends Genet.*, **31**, 509–515.
3. Knipe,D.M. (2015) Nuclear sensing of viral DNA, epigenetic regulation of herpes simplex virus infection, and innate immunity. *Virology*, **479–480**, 153–159.
4. Dempsey,A. and Bowie,A.G. (2015) Innate immune recognition of DNA: a recent history. *Virology*, **479–480**, 146–152.
5. Xia,P., Wang,S., Ye,B., Du,Y., Huang,G., Zhu,P. and Fan,Z. (2015) Sox2 functions as a sequence-specific DNA sensor in neutrophils to initiate innate immunity against microbial infection. *Nat. Immunol.*, **16**, 366–375.
6. Song,J., Li,Z., Tong,X., Chen,C., Chen,M., Meng,G., Chen,P., Li,C., Xin,Y., Gai,T. *et al.* (2015) Genome-wide identification and characterization of Fox genes in the silkworm, Bombyx mori. *Funct. Integr. Genomics*, **15**, 511–522.
7. Seo,E., Choi,D. and Choi. (2015) Functional studies of transcription factors involved in plant defenses in the genomics era. *Brief. Funct. Genomics*, **14**, 260–267.

8. Sanchez-Romero,M.A., Cota,I. and Casadesus,J. (2015) DNA methylation in bacteria: from the methyl group to the methylome. *Curr. Opin. Microbiol.*, **25**, 9–16.
9. Meng,H., Cao,Y., Qin,J., Song,X., Zhang,Q., Shi,Y. and Cao,L. (2015) DNA methylation, its mediators and genome integrity. *Int. J. Biol. Sci.*, **11**, 604–617.
10. Mayran,A., Pelletier,A. and Drouin,J. (2015) Pax factors in transcription and epigenetic remodelling. *Semin. Cell Dev. Biol.*, **44**, 135–144.
11. Marchal,C. and Miotto,B. (2015) Emerging concept in DNA methylation: role of transcription factors in shaping DNA methylation patterns. *J. Cell. Physiol.*, **230**, 743–751.
12. Eelen,G., Verlinden,L., Maes,C., Beullens,I., Gysemans,C., Paik,J.H., DePinho,R.A., Bouillon,R., Carmeliet,G. and Verstuyf,A. (2015) Forkhead box O transcription factors in chondrocytes regulate endochondral bone formation. *J. Steroid Biochem. Mol. Biol.*, **164**, 337–343.
13. Woo,S.R., Fuertes,M.B., Corrales,L., Spranger,S., Furdyna,M.J., Leung,M.Y., Duggan,R., Wang,Y., Barber,G.N., Fitzgerald,K.A. *et al.* (2014) STING-dependent cytosolic DNA sensing mediates innate immune recognition of immunogenic tumors. *Immunity*, **41**, 830–842.
14. Atianand,M.K. and Fitzgerald,K.A. (2013) Molecular basis of DNA recognition in the immune system. *J. Immunol.*, **190**, 1911–1918.
15. Gao,M. and Skolnick,J. (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic Acids Res.*, **36**, 3978–3992.
16. Ahmad,S., Gromiha,M.M. and Sarai,A. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
17. Ahmad,S., Keskin,O., Sarai,A. and Nussinov,R. (2008) Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Res.*, **36**, 5922–5932.
18. Andrabi,M., Mizuguchi,K. and Ahmad,S. (2014) Conformational changes in DNA-binding proteins: relationships with precomplex features and contributions to specificity and stability. *Proteins*, **82**, 841–857.
19. Andrabi,M., Mizuguchi,K., Sarai,A. and Ahmad,S. (2009) Prediction of mono- and di-nucleotide-specific DNA-binding sites in proteins using neural networks. *BMC Struct. Biol.*, **9**, 30.
20. Andrabi,M., Mizuguchi,K., Sarai,A. and Ahmad,S. (2008) Benchmarking and analysis of DNA-binding site prediction using machine learning methods. In: *Proceedings of IEEE International Joint Conference Neural Networks, June 1–6*. Hong Kong, Vol. **NN0554**, pp. 1746–1750.
21. Zhou,W. and Yan,H. (2011) Prediction of DNA-binding protein based on statistical and geometric features and support vector machines. *Proteome Sci.*, **9**(Suppl. 1), S1.
22. Si,J., Zhang,Z., Lin,B., Schroeder,M. and Huang,B. (2011) MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. *BMC Syst. Biol.*, **5**(Suppl. 1), S7.
23. Ahmad,S. (2011) In: Kihara,D (ed). *Protein Function Prediction for Omics Era*. Springer Netherlands, pp. 165–182.
24. Huang,Y.F., Huang,C.C., Liu,Y.C., Oyang,Y.J. and Huang,C.K. (2009) DNA-binding residues and binding mode prediction with binding-mechanism concerned models. *BMC Genomics*, **10**(Suppl. 3), S23.
25. Li,T., Li,Q.Z., Liu,S., Fan,G.L., Zuo,Y.C. and Peng,Y. (2013) PreDNA: accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information. *Bioinformatics*, **29**, 678–685.
26. Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–W248.
27. Wu,J., Liu,H., Duan,X., Ding,Y., Wu,H., Bai,Y. and Sun,X. (2009) Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics*, **25**, 30–35.
28. Miao,Z. and Westhof,E. (2015) Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. *Nucleic Acids Res.*, **43**, 5340–5351.
29. Zhang,Y., Xu,J., Zheng,W., Zhang,C., Qiu,X., Chen,K. and Ruan,J. (2014) newDNA-Prot: Prediction of DNA-binding proteins by employing support vector machine and a comprehensive sequence representation. *Comput. Biol. Chem.*, **52**, 51–59.
30. Yan,C. and Wang,Y. (2014) A graph kernel method for DNA-binding site prediction. *BMC Syst. Biol.*, **8**(Suppl. 4), S10.
31. Wang,W., Liu,J., Xiong,Y., Zhu,L. and Zhou,X. (2014) Analysis and classification of DNA-binding sites in single-stranded and double-stranded DNA-binding proteins using protein information. *IET Syst. Biol.*, **8**, 176–183.
32. Park,B., Im,J., Tuvshinjargal,N., Lee,W. and Han,K. (2014) Sequence-based prediction of protein-binding sites in DNA: comparative study of two SVM models. *Comput. Methods Programs Biomed.*, **117**, 158–167.
33. Niu,X.H., Hu,X.H., Shi,F. and Xia,J.B. (2014) Predicting DNA binding proteins using support vector machine with hybrid fractal features. *J. Theor. Biol.*, **343**, 186–192.
34. Lou,W., Wang,X., Chen,F., Chen,Y., Jiang,B. and Zhang,H. (2014) Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes. *PLoS One*, **9**, e86703.
35. Liu,B., Xu,J., Lan,X., Xu,R., Zhou,J., Wang,X. and Chou,K.C. (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**, e106691.
36. Levinson,M. and Zhou,Q. (2014) A penalized Bayesian approach to predicting sparse protein-DNA binding landscapes. *Bioinformatics*, **30**, 636–643.
37. Zou,C., Gong,J. and Li,H. (2013) An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis. *BMC Bioinformatics*, **14**, 90.
38. Ma,X., Wu,J. and Xue,X. (2013) Identification of DNA-binding proteins using support vector machine with sequence information. *Comput. Math. Methods Med.*, **2013**, 524502.
39. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
40. Firoz,A., Malik,A., Joplin,K.H., Ahmad,Z., Jha,V. and Ahmad,S. (2011) Residue propensities, discrimination and binding site prediction of adenine and guanine phosphates. *BMC Biochem.*, **12**, 20.
41. Malik,A. and Ahmad,S. (2007) Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. *BMC Struct. Biol.*, **7**, 1.
42. Fernandez,M., Kumagai,Y., Standley,D.M., Sarai,A., Mizuguchi,K. and Ahmad,S. (2011) Prediction of dinucleotide-specific RNA-binding sites in proteins. *BMC Bioinformatics*, **12**(Suppl. 13), S5.
43. Ahmad,S. and Mizuguchi,K. (2011) Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. *PLoS One*, **6**, e29104.
44. Dozmorov,M.G., Giles,C.B. and Wren,J.D. (2011) Predicting gene ontology from a global meta-analysis of 1-color microarray experiments. *BMC Bioinformatics*, **12**(Suppl. 10), S14.
45. Blaby-Haas,C.E. and de Crecy-Lagard,V. (2011) Mining high-throughput experimental data to link gene and function. *Trends Biotechnol.*, **29**, 174–182.
46. Wren,J.D. (2009) A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. *Bioinformatics*, **25**, 1694–1701.
47. Cheng,W.C., Tsai,M.L., Chang,C.W., Huang,C.L., Chen,C.R., Shu,W.Y., Lee,Y.S., Wang,T.H., Hong,J.H., Li,C.Y. *et al.* (2010) Microarray meta-analysis database (M(2)DB): a uniformly pre-processed, quality controlled, and manually curated human clinical microarray database. *BMC Bioinformatics*, **11**, 421.
48. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
49. Andrabi,M., Mizuguchi,K., Sarai,A. and Ahmad,S. (2009) Prediction of mono- and di-nucleotide-specific DNA-binding sites in proteins using neural networks. *BMC Struct. Biol.*, **9**, 30.
50. Ahmad,S., Gromiha,M.M. and Sarai,A. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based

on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
51. Nagarajan,R., Ahmad,S. and Gromiha,M.M. (2013) Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins. *Nucleic Acids Res.*, **41**, 7606–7614.
52. Xiong,Y., Liu,J. and Wei,D.Q. (2011) An accurate feature-based method for identifying DNA-binding residues on protein surfaces. *Proteins*, **79**, 509–517.
53. Wang,L., Huang,C., Yang,M.Q. and Yang,J.Y. (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.*, **4**(Suppl. 1), S3.
54. Cai,Y., He,Z., Shi,X., Kong,X., Gu,L. and Xie,L. (2010) A novel sequence-based method of predicting protein DNA-binding residues, using a machine learning approach. *Mol. Cells*, **30**, 99–105.
55. Ahmad,S. and Sarai,A. (2004) Moment-based prediction of DNA-binding proteins. *J. Mol. Biol.*, **341**, 65–71.
56. Kuhn,M. (2008) Caret package. *J. Stat. Softw.*, **28**, http://download.nextag.com/cran/web/packages/caret/caret.pdf.
57. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
58. Kind,T. and Fiehn,O. (2009) What are the obstacles for an integrated system for comprehensive interpretation of cross-platform metabolic profile data? *Bioanalysis*, **1**, 1511–1514.
59. Lin,S.M., Du,P., Huber,W. and Kibbe,W.A. (2008) Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res.*, **36**, e11.
60. Lin,X. and Chen,X.W. (2013) Heterogeneous data integration by tree-augmented naive Bayes for protein-protein interactions prediction. *Proteomics*, **13**, 261–268.
61. Chen,Y.A., Tripathi,L.P. and Mizuguchi,K. (2011) TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery. *PloS One*, **6**, e17844.
62. Chen,Y.-A., Tripathi,L.P. and Mizuguchi,K. (2016) An integrative data analysis platform for gene set analysis and knowledge discovery in a data warehouse framework. *Database*, **2016**, baw009.
63. Ma,W., Noble,W.S. and Bailey,T.L. (2014) Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nat. Protoc.*, **9**, 1428–1450.
64. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.
65. Yan,J. and Kurgan,L. (2017) DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res.*, **45**, e84.
66. Peng,Z. and Kurgan,L. (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.*, **43**, e121.
67. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 2444–2448.
68. Zhao,H., Wang,J., Zhou,Y. and Yang,Y. (2014) Predicting DNA-Binding Proteins and Binding Residues by Complex Structure Prediction and Application to Human Proteome. *PLoS ONE*, **9**, e96694.
69. Karas,V.O., Westerlaken,I. and Meyer,A.S. (2015) The DNA-binding protein from starved cells (Dps) utilizes dual functions to defend cells against multiple stresses. *J. Bacteriol.*, **197**, 3206–3215.
70. Bolduc,N., Hake,S. and Jackson,D. (2008) Dual functions of the KNOTTED1 homeodomain: sequence-specific DNA binding and regulation of cell-to-cell transport. *Sci. Signal.*, **1**, pe28.