UNIVERSITY OF CALIFORNIA SAN DIEGO

The Promoter Sequence Basis for Transcription Regulatory Activity

A thesis submitted in partial satisfaction of the requirements

for the Master of Science degree

in

Bioengineering

By

Sizhe Qiu

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Jeff Hasty
Professor Milton Saier

2021

The Thesis of Sizhe Qiu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank Professor Bernhard Ø. Palsson and Dr. Daniel Zielinski for providing the opportunity and wonderfully collaborative lab environment to complete this study.

I would also like to thank the members of the Systems Biology Research Group for their collaboration and guidance especially Dr. Anand Sastry, Cam Lamoureux, Dr.Gao Ye and Patrick Phaneuf.

ABSTRACT OF THE THESIS

The Promoter Sequence Basis for Transcription Regulatory Activity

By

Sizhe Qiu

Master of Science in Bioengineering

University of California San Diego, 2021

Professor Bernhard Ø. Palsson, Chair

The transcriptional regulatory network (TRN) of E. coli MG1655 contains thousands of regulatory interactions between transcription factors, sigma factors and promoter sequences of genes. The membership of a regulon, in other words whether a gene is regulated by a transcription factor, has a basis in the gene's promoter. To determine the TRN, there are two primary families of methods: bottom-up identification of binding sites via methods such as ChIP-seq experiments, and top-down inference of the TRN via expression analysis methods such as

independent component analysis (ICA). In this work, the promoter sequence of each gene was utilized to predict regulons with machine learning. Certain promoter features, such as TF binding site motif scores, DNA shape features, and sigma factor-related features were found to be essential to predict whether a gene is regulated by particular transcription factors. ICA and ChIP TRNs were compared to investigate factors underlying their difference, and two case studies were carried out. An FNR case study showed that ICA regulons fractionate a large ChIP regulon into several regulation types. An ArcA case study demonstrated that the ICA TRN captures diversity in binding sites' architecture. In general, through comparison, ICA TRN extracts genes of strong regulation activity from ChIP TRN. We then expanded this analysis to understand differences in the regulons of multiple strains of E. coli. A pan-regulon of Fur was reconstructed with unique, accessory and core regulons annotated. We found that genes in the core regulon have stronger regulation activity than genes in the unique regulon. Additionally, it was also found that the motif score and helix twist of DNA sequence were both significant indicators of Fur ChIP-exo peak heights, represented by S/N ratios. This study was a meaningful application of machine learning on biological problems that probed biophysical factors underlying omics data, giving directions for the genome design in synthetic biology aiming to control the phenotype by TRN tuning.

INTRODUCTION

Bacterial transcription is the process of making a copy of the bacterial DNA sequence into mRNA by RNA polymerase. The understanding and control of bacterial transcription has become an increasingly important topic with more and more studies done to explore its regulation behavior, due to great effectiveness of making products with bacterial genetic materials. Though every step between gene and function is exploited to implement the control of transcription, for reasons of economy, the key step to regulate is transcription initiation, which is a highly regulated process [1]. For transcription initiation, RNA polymerase recognizes the promoter sequence upstream to coding region with the polymerase sigma factor binding to -10 and -35 boxes, C-terminal domain binding to an AT rich upstream element. Then, the open-complex of RNA polymerase and DNA sequence forms.

Apart from sigma factors' contribution to initiation, transcription factors bind to promoter sequence, mostly upstream to transcription start sites, and activate or repress transcription. In E. coli, 115 TFs, 7 sigma factors are identified with a list of associated regulons encompassing sets of genes regulated by the same regulatory protein [2]. Regulons together constitute the transcriptional regulatory network (TRN) with thousands of regulatory interactions between proteins and promoter sequences discovered. To understand the mechanism underlying the formation of TRN, one should investigate the promoter sequence and look for important sequence features impacting the TRN.

ChIP (chromatin immunoprecipitation) is usually used to determine transcription factor binding sites and the TRN experimentally. Bacterial cells are treated with formaldehyde to crosslink proteins and DNA. DNA fragments enriched with the protein of interest are immunoprecipitated and then sequenced to identify genome wide sites [3]. The resulting reads

would be aligned to the genome by computational tools like Bowtie, and ChIP-seq peaks, the sites of enrichment in ChIP-seq data are then obtained. With sites of high enrichment determined on genome sequence, ChIP regulons of specific regulators are constructed, and ChIP TRN is then determined.

Apart from the ChIP experiment, independent component analysis (ICA) serves as another approach to determining the TRN by decomposing the data matrix of transcriptome. ICA is an unsupervised statistical learning method used to find a linear representation of non-Gaussian data so that the components are statistically independent [4]. Similar to its application to separate mixed sound signals, it could also be used to analyze transcription levels modulated by TFs and sigma factors. In Dr. Anand V. Sastry's paper, The Escherichia coli transcriptome mostly consists of independently regulated modules, 92 stable independent components were obtained from a diverse compendium of over 250 high-quality Escherichia coli RNA-seq datasets [5]. In order to obtain the set of significant genes in each component, genes with largest absolute value coefficients were iteratively removed and D'Agostino K2 test statistic was computed. Once the statistic below the cutoff, the set of removed genes were determined as significant, forming I-modulon, which is the ICA regulon. Then, ICA regulons were annotated with single or multiple regulators based on ChIP regulon enrichment.

For TRN determination, ChIP TRN is determined from binding sites on genome sequence and corresponding genes, while ICA TRN is determined as sets of significant genes from gene expression modulated by mixed regulatory activities. In contrast, ChIP TRN is determined structurally by analyzing DNA fragments enriched with bound proteins, while ICA TRN is determined functionally by decomposing the resulting transcription signal. Also, ChIP regulons usually have single TFs associated, but ICA regulons could represent co-regulation of

multiple TFs. ChIP regulons have motif PSSMs (position specific scoring matrix) of binding sites, while ICA regulons only have memberships identified. Comparison of two types of TRNs would reveal the sequence basis of the difference, leading to better interpretation of ICA and ChIP TRNs.

The regulon of the same regulator would vary across strains, and one case is Fur (Ferric uptake regulator) that regulates iron homeostasis. Though Fur's protein sequence is conserved across strains, it plays distinct roles in different E. coli strains [6]. Using ChIP-exo experiments, Dr. Gao Ye performed a comprehensive study of Fur binding sites in multiple E. coli strains. From the pan-regulon reconstructed, he discovered that 18% genes belong to core regulon, 42% to accessory regulon and 40% to unique regulon in each strain [6]. The ChIP-exo experiments provide us S/N ratios of binding sites in each strain, which is signal to noise ratios representing regulator binding strength of each site. To get a more in depth understanding of differential regulation activity across strains, it is necessary to probe what promoter sequence features influence S/N ratios of binding sites most in multiple strains.

In this study, the aim was to compute promoter sequence features with genome information curated and use machine learning tools to investigate three topics: genome sequence features that determine ICA TRN, the cause of difference between ChIP and ICA TRNs, and sequence basis of regulon's variation across strains. The goal was to get a mechanistic interpretation of transcriptional regulation activity by figuring out specific biophysical factors related, which would eventually provide better knowledge for genome design to control bacteria's transcription.

METHODS

2.1 Bitome: the tool used to digitize genome information curated

The model organism in this study was Escherichia coli, and genome information was mainly curated from RegulonDB, NCBI and Prokka. The Bitome tool developed by Cam Lamourex was utilized to digitize and organize genome information. With information tables loaded, it can build a bitome object that maps multiple classes of genomic objects onto genome sequence, such as coding sequences, transcription units, operons, transcription factor binding sites and so on [7]. The tool enables researchers to locate genomic objects and specific regions efficiently, assisting sequence feature computation. For example, a bitome object contains transcription start sites associated with transcription units, and hence, one just needs to iterate all transcription start sites to compute promoter sequence features near them and determine relative positions to transcription start sites, upstream or downstream.

2.2 Motif score of promoter sequence

In general, the motif score of a specific transcription factor or sigma factor is computed by searching for the segment of highest probability to be the binding site. As PSSM's coefficients are log-odds of probability, summation of log-odds at all sites could represent the probability of motif existence. The formula is $motif\_score = \sum_{i=0}^{k} M[\,i\,][\,S[\,i\,]\,]$, and M is PSSM, k is motif length, S is the segment, i is the index. The algorithm performs 1-D dynamic programming that iteratively computes the motif score of consecutive equal-length segments (length = motif length), and the segment of highest score is the matched motif box. Then, the highest score is output as the motif score of the promoter sequence.

For transcription factors, the search range was -150bp and +30bp to the transcription start site, and in this study, PSSMs were obtained from RegulonDB's TF PWMs Browser page, mainly determine experimentally by ChIP seq experiments [8]. Unlike TF motifs, motifs of sigma factors are two binding boxes with a gap of varying length around 17bp, -10 box (pribnow box) and -35 box. PSSMs of -10 and -35 boxes of sigma 70/32/38/24/54/28 were generated from annotated sigma factor binding sites in RegulonDB and consensus sequences in Transcription regulation in prokaryotes by Rolf Wagner [9]. Other sigma factors were not included due to limited annotation. The search range was 20bp upstream for -10 box, 40bp for -35 box. In addition to motif scores of -10/-35 boxes on promoter sequences, an AT-rich 7bp segment in the spacer between -10 and -35 boxes was matched and the highest AT ratio was output to account for the extended -10 box [10]. Sigma factor related features were motif scores of -10/-35 boxes, the distance from -10/-35 boxes to the transcription start site, spacer length, AT ratio of the spacer, and AT ratio of the extended -10 box. To reduce feature dimensions, a hyperplane decision surface, dividing the sigmulon (genes regulated by the sigma factor) and other genes, was computed using LDA (linear discriminant analysis), and features were projected to it.

Some motifs in RegulonDB fail to capture the whole binding site consensus sequence, or other motifs also important to regulation activity are missed. For example, Fur binding site motif in RegulonDB fails to include the whole 19bp length inverted complementary repeats structure [11]. Motif discovery in ICA regulons using MEME suite provides hidden motifs missed by RegulonDB. MEME is a toolbox used to discover novel motifs in collections of unaligned sequences [12]. Those ICA regulon motifs were used to supplement information missed by ChIP experiments.

Another possible edge case was that the regulator binding site's architecture has great diversity, which could not be characterized by a simple motif. For instance, 42% of ArcA binding sites are 2 direct repeats of 10bp segments, 41% are 3 direct repeats, 15% are 4 direct repeats and 5% are 5 direct repeats [13]. And hence, DR motifs were used to build the classifier to predict ArcA ICA regulon membership, the performance was compared to the original classifier with ArcA ChIP motif features.

2.3 DNA shape feature computation

DNA shape of TF binding sites could provide structural information on the interaction between regulators and sequence, as regulator binding affinity is strongly related to the structure of the binding site. For example, minor groove width indicates the size of the binding pocket. In this study, 13 types of shape features were computed, 6 inter-base pair shapes: Roll, HelT (Helix Twist), Shift, Slide, Rise and Tilt, and 7 intra-base pair shapes: MGW (Minor Groove Width), ProT (Propeller Twist), Buckle, Shear, Stretch, Stagger, Buckle and Opening.

The pentamer query table integrated with a sliding-pentamer window was used to compute shape vectors for all matched TF binding sites. The pentamer table was obtained from the github repository of DNAShapeR, a DNA shape predictor in R language [14]. For predicting intra-base pair parameters, each sliding step assigned a shape prediction for the central base pair. For predicting inter-base pair parameters, each sliding step assigned a shape prediction for two central base-pair steps. The overlapping values arising from two adjacent pentamers at the same nucleotide position were averaged. The sliding-window approach would result in a shape vector [15]. Maximum, minimum, range and mean values of the shape vector were computed as

features. To reduce dimensions of shape features, LDA was applied to project shape features to the hyperplane decision surface, which divides the ChIP regulon and other genes.

2.4 Logistic regression classification on ICA regulon membership and assessment of model performance

Logistic regression classifier was chosen in this study because it is a more interpretable simplistic linear model compared to Random Forest or Neural Network, and the degree of overfitting is usually smaller. It was implemented using scikit-learn, a free software machine learning library for python [16]. The target labels were binarized ICA regulon memberships. Two feature matrices were trialed, only ChIP motif scores and all features. Model performance was assessed by 5-fold cross validation and area under curve of ROC (receiver operating characteristic) showing true positive rate versus false positive rate across all possible classification thresholds. A cutoff = 0.8 was used to indicate good performance. For optimal classification performance, hyperparameters of LR classifier were randomly searched, such as l1/l2 penalty, maximum iteration, and the combination with highest area under curve of ROC would be selected.

2.5 Promoter prediction and pan-regulon reconstruction

Though genomic information of E. coli MG1655 is well annotated, other strains,such as W3110, have limited genomic objects annotated, especially transcription units with transcription start sites determined that indicate location of promoters. Therefore, predicting promoters' locations was necessary for multi-strain study of E. coli. Assuming genome sequences are mostly

conserved across strains, I aligned sequences of transcription units in MG1655 strain to other strains' genome sequences using BLAST, which is a tool enabling a researcher to query a subject sequence with a database of sequences [17]. Then, alignments without left and right ends exactly matched were filtered out, only exact matches were used to build transcription unit objects mapped onto genome sequence. Subsequently, based on strand orientation, transcription start sites were determined as left end for forward strand, right end for reverse strand.

After promoters were predicted for each strain, Fur ChIP-exo peaks with S/N ratios were mapped to promoter sequences, and transcription units regulated by Fur in each strain were determined. Since predicted transcription units were all exact matches, genes in transcription units were known, therefore genes regulated by Fur were also determined. The Fur pan-regulons of genes and transcription units were then reconstructed under bitome framework for six strains of E. coli: MG1655, KO11FL, CFT017, W, W3110 and BL21, across three phylogroups: A, B2 and F. Unique, accessory and core regulons were annotated.

2.6 Logistic regression classification on high or low S/N ratios of Fur ChIP-exo sites in multiple strains

The LR classifier was implemented in the same way as in section 2.4, and performance was also assessed using the area under curve of ROC. A cutoff = 10 (2.3 in log-scale) separated each strain's bimodal distribution of Fur ChIP exo sites' S/N ratios into high and low values, which was binarized into target labels for classification. Features used included all features computed in section 2.2 and 2.3, the same as LR classification on ICA regulon membership. Hyperparameter optimization was done in the same manner as in section 2.4.

RESULTS

3.1 Essential features to determine the ICA TRN

Initial classification models of ICA regulon memberships only used ChIP motif scores as features, resulting in large prediction errors for some models shown in Figure 1. Most models' AUC ROC scores were around 0.7. Therefore, ChIP motif scores are not sufficient to predict the ICA TRN.

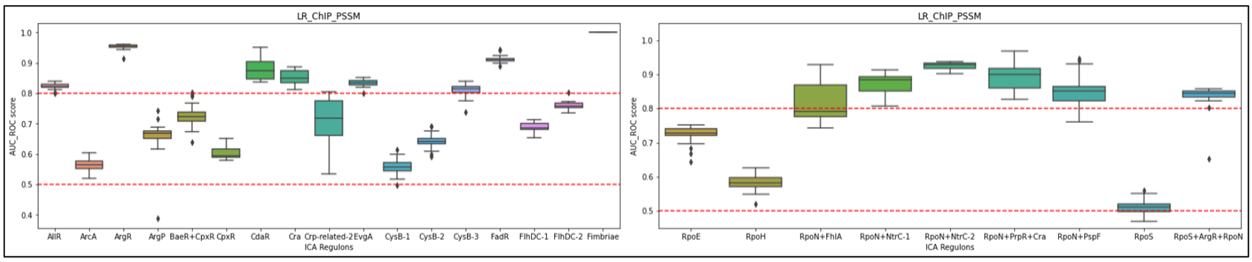

Figure 1. Initial classification model performance with only ChIP motif scores. Some AUC ROC scores are under 0.8, indicating low prediction accuracy. Models of RpoH (Sigma32) and RpoS (Sigma38) had low performance, close to 0.5.

To account for sigma factor associated ICA regulons, sigma factor binding sites were matched and scored. Sigma factor motifs were matched accurately (Figure 2A). At TF binding sites, DNA shapes show characteristic patterns, compared to those of random sites, suggesting that shape features would assist improvement of model performance (Figure 2C). Sigma factor related features and shape features were processed by LDA, shown in Figure 2CD. They both have significant distinguishing distributions between the sigmulon or ChIP regulon and other genes.

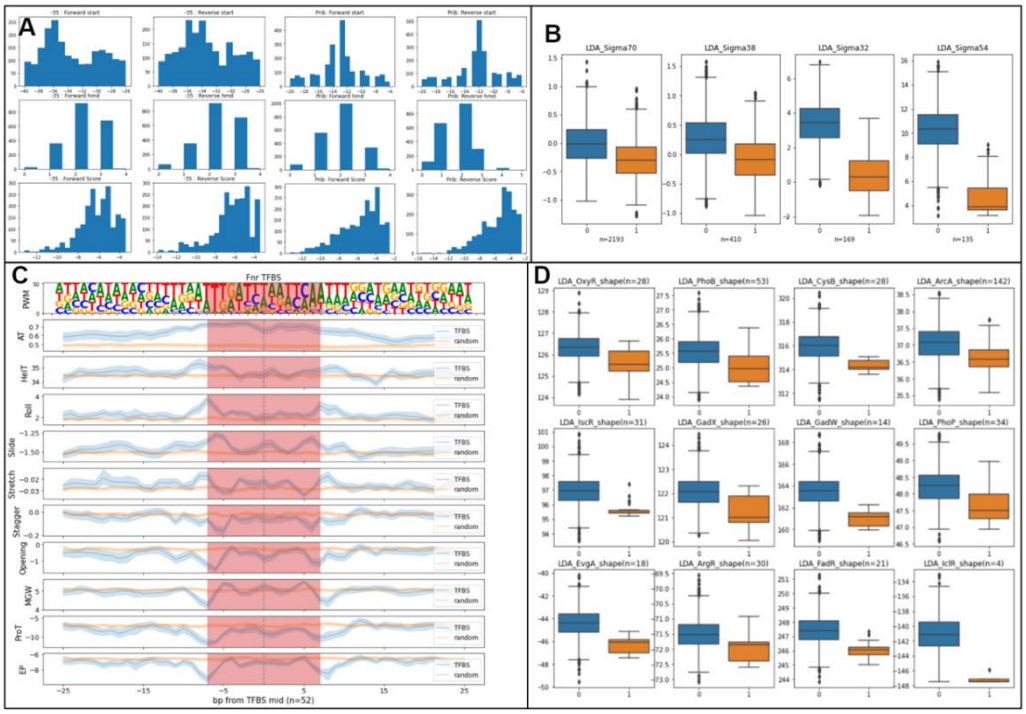

Figure 2. Display of additional features to assist the prediction. (A) Sigma factor binding box match result. Matched positions were consistent with theoretical positions. (B) Distribution of sigma factor related features processed by LDA (C) Profile of DNA shape features at TF binding sites and random sites. (D) Distribution of DNA shape features processed by LDA.

With all features: motif scores of TFs and sigma factors, DNA shape and genome organization information such as strand direction, most classification models' performance improved that they were above the 0.8 cutoff, but for some models, such as ArcA, still remained inaccurate (Figure 3). For sigma factor associated ICA regulons, the classification model of RpoH (Sigma32) had a great accuracy improvement, the sigma factor related features of RpoH, in Figure 2B, separate two groups well, in or not in sigmulon. However, the model of RpoS (Sigma38) still had bad performance. Because RpoS recognizes the same promoters as RpoD (Sigma 70), which has the largest sigmulon, the difficulty of finding distinguishing features results in large prediction error for RpoS ICA regulon membership [9].

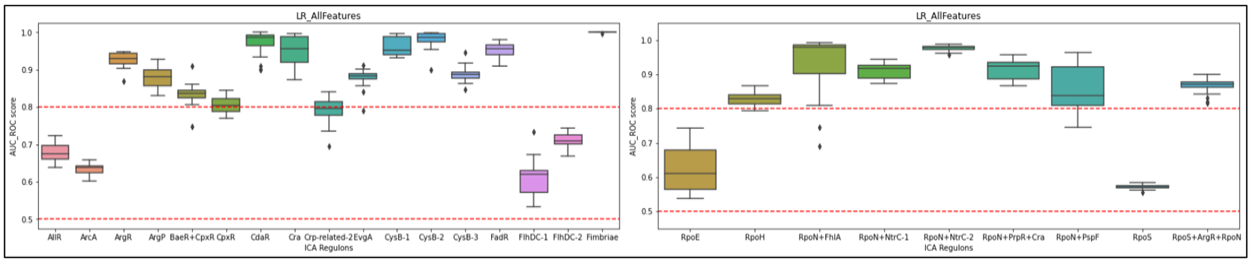

Figure 3. Assessment of classification model with all features included. Most models' performances improved, especially for classification models of CysB associated ICA regulons. The classification model of RpoH improved, while that of RpoS remained inaccurate.

In summary, there are currently 53 models above 0.8 cutoff, 30 out of 39 remaining models cannot be built due to lack of data, such as binding site motif of associated regulator(s). The TF motif score is the most significant indicator of regulation activity, but not sufficient to predict ICA TRN, and prediction accuracy can be largely improved by inclusion of sigma factor related features, DNA shape features etc. More analysis of promoter sequence needs to be done to account for special cases for further improvement of prediction accuracy.

3.2 Comparison between TRNs from ChIP experiment and ICA

Though ICA and ChIP regulons often have large overlaps, there are distinctions between them for certain regulators, such as Fnr and ArcA. There are 5 Fnr associated ICA regulons: Fnr+IHF+gcvB, Fnr+NarL, Fnr+NarLP and Fnr -1 and 2, which are still not well annotated. From the venn diagram in Figure 4A, ICA TRN identifies 3 regulation types of Fnr: co-regulation with NarL or NarP, co-regulation with IHF (integration host factor), and the classification result and feature importances in Figure 4CD also support the regulation type identification. Compared to Fnr ChIP motif, two ICA regulon motifs appear to be better indicators of ICA regulon memberships for Fnr+IHF+gcvB, Fnr+NarL and Fnr+NarLP, shown in Figure 4E. And the sequence basis of NarL/P's co-regulation with Fnr is verified by high distributions of motif scores in Fnr+NarL and Fnr+NarP ICA regulons. In short, ICA TRN's identification of Fnr's regulation types is well supported by the study on promoter sequence basis.

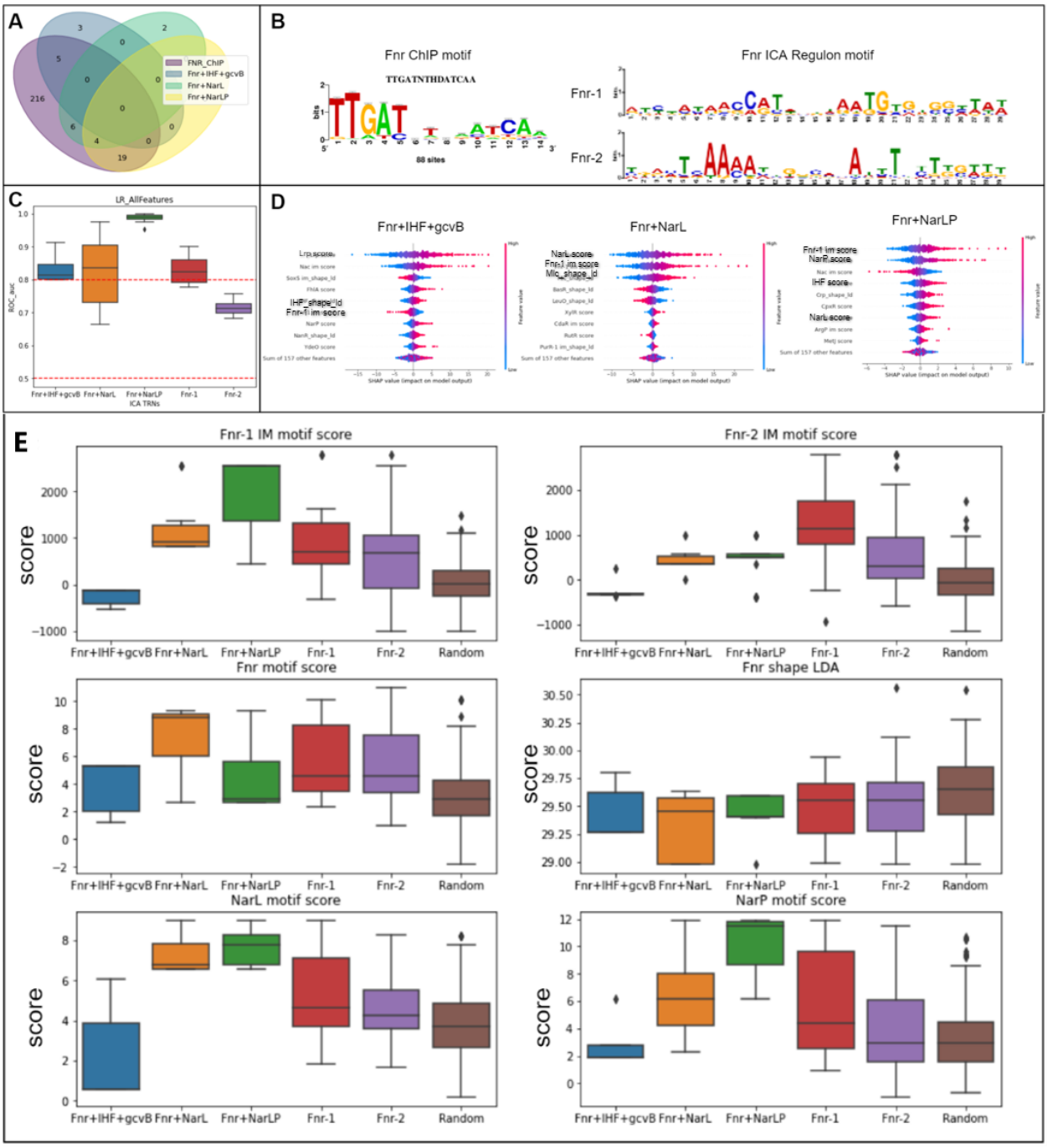

Figure 4. Fnr case study result. (A) The venn diagram displaying relationships of 3 well annotated Fnr ICA regulons and ChIP regulon. Three ICA regulons are fractions extracted from the large ChIP regulon. (B) Fnr's ChIP motif and Fnr-1 and 2 ICA regulon motifs. (C) Assessment of classification models for 5 Fnr associated ICA regulons. (D) Feature importances of classification models for Fnr+IHF+gcvB, Fnr+NarL and Fnr+NarLP. (E) Distributions of features across ICA regulons: Fnr-1 and Fnr-2 ICA regulon motif scores, Fnr ChIP motif score and shape feature,NarL and NarP motif scores.

The binding site of ArcA has great diversity, and hence classification models using multiple direct repeats motifs and ChIP motif separately were compared. DR motif scores are better indicators of ArcA ICA/ChIP regulon memberships than the ChIP motif, as the classification model had better performance (Figure 5B). DR motifs characterize the architecture of binding sites better: the information content and number of DR elements. The lower distributions of 3DR and 4DR motif scores in ArcA ChIP regulon suggest that it fails to collect genes with promoters containing extended DR elements, and ICA regulon is a good supplement.

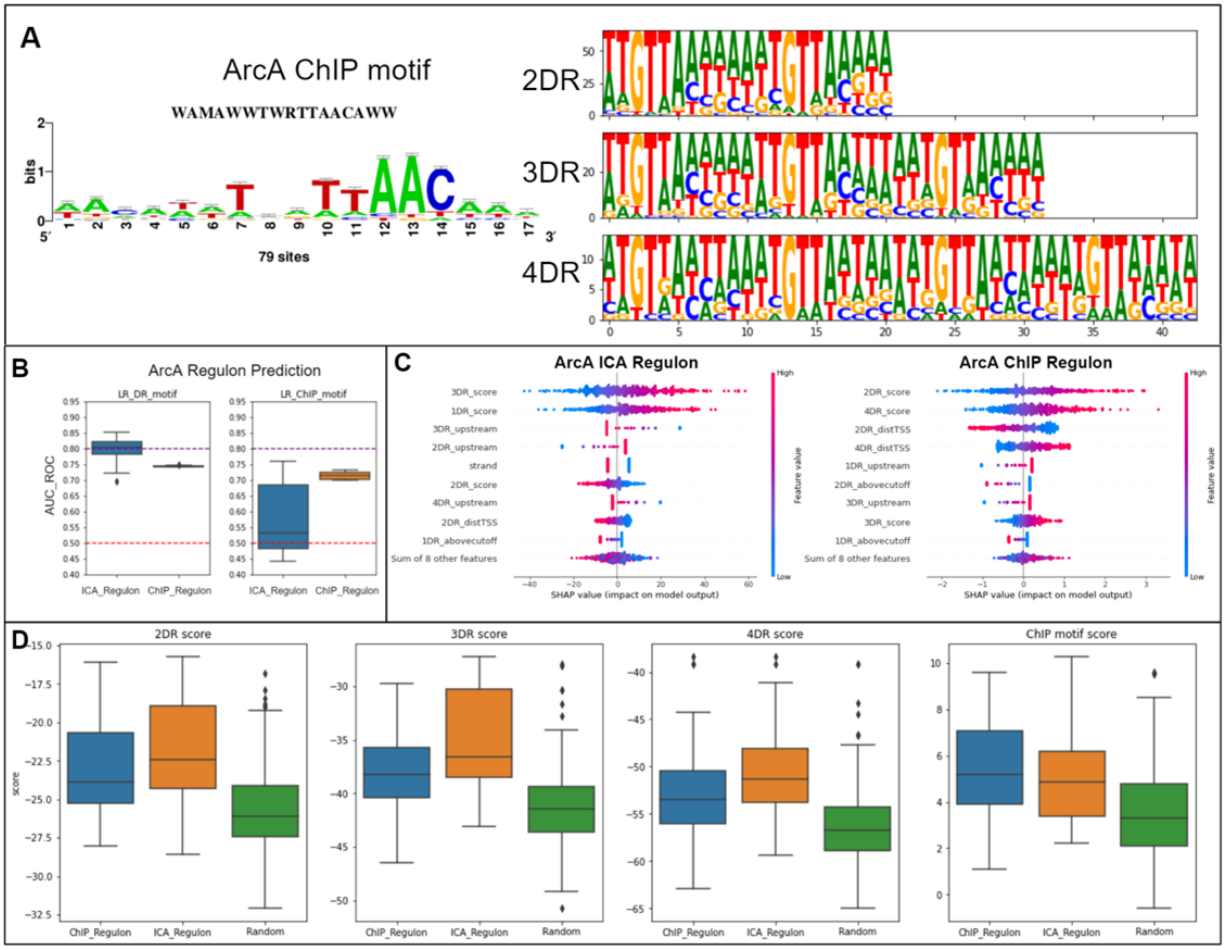

Figure 5. ArcA case study result. (A) ArcA's ChIP motif and 2,3,4 direct repeats motifs. (B) Assessment of classification models for AcrA ICA and ChIP regulons using ChIP motif and DR motifs. (C) Feature importances of classification models using DR motifs. The most important feature in predicting ArcA ICA and ChIP regulons are 3DR and 2DR motif scores. (D) Distributions of DR and ChIP motif scores across regulons.

From case studies of Fnr and ArcA, it was observed that the distribution of motif scores tends to be higher in ICA regulons than those of ChIP regulons. In order to check the observation is valid, 9 single TF associated ICA regulons were compared to ChIP regulons. The result in Figure 6 demonstrates that ICA TRN extracts genes of strong regulation activity from ChIP TRN, which can be interpreted from higher distributions of binding site strength characterized by motif scores in ICA regulons.

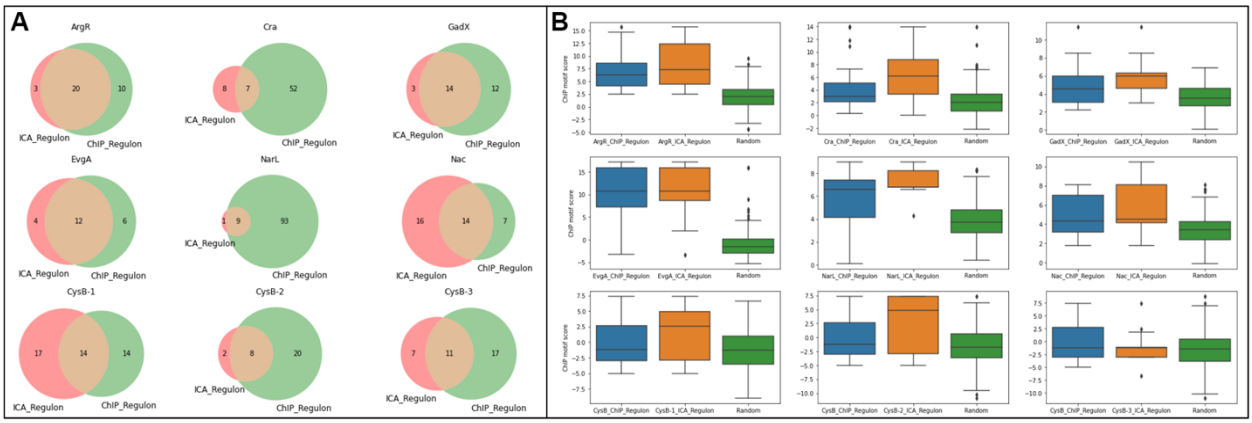

Figure 6. Comparison between ICA and ChIP regulons. (A) Venn diagrams showing overlaps between two types of regulons. (B) Comparison of motif scores between 2 types of regulons.

In conclusion, ICA TRN is different from ChIP TRN because ICA TRN could discover information missed by ChIP experiments, such as categorized regulation types, diversity in binding site architecture and genes with strong regulation activity.

3.3 Sequence basis of TRN's variation across strains

To explore TRN's variation across multiple E. coli strains, Fur pan-regulons of genes and transcription units for 6 strains across 3 phylogroups were reconstructed under bitome framework, displayed in Figure 7A. In Figure 7B, S/N ratios of genes in core regulon were compared, W and KO11FL in F phylogeny group have higher S/N ratios compared to MG1655, while CFT073 in B2 group has relatively low S/N ratios, indicating lower regulation activity. Strains in the same phylogroup tend to have closer regulation activity.

Genes with ChIP exo peaks on promoters were grouped into ones with high and low S/N ratios, shown in Figure 7C. From Figure 7DE, we could see that while genes of low S/N ratios overlap little with Fur ChIP/ICA regulons, ICA TRN determined with MG1655's transcription levels is robust at capturing genes of high S/N ratios in multiple strains, which is consistent with the finding in the previous section that ICA TRN contains genes of strong regulation activity.

Only genes in core regulon are regulated in each strain, unique and accessory regulons vary across strains. A tendency showing Fur regulation activity: core > accessory > unique can be observed in Figure 7F, and the similar tendency of motif scores suggests the major resulting factor of regulation activity variation across core, unique and accessory regulons is binding site strength. Conserved promoters across strains tend to have high binding site strength, and thus high regulation activity.

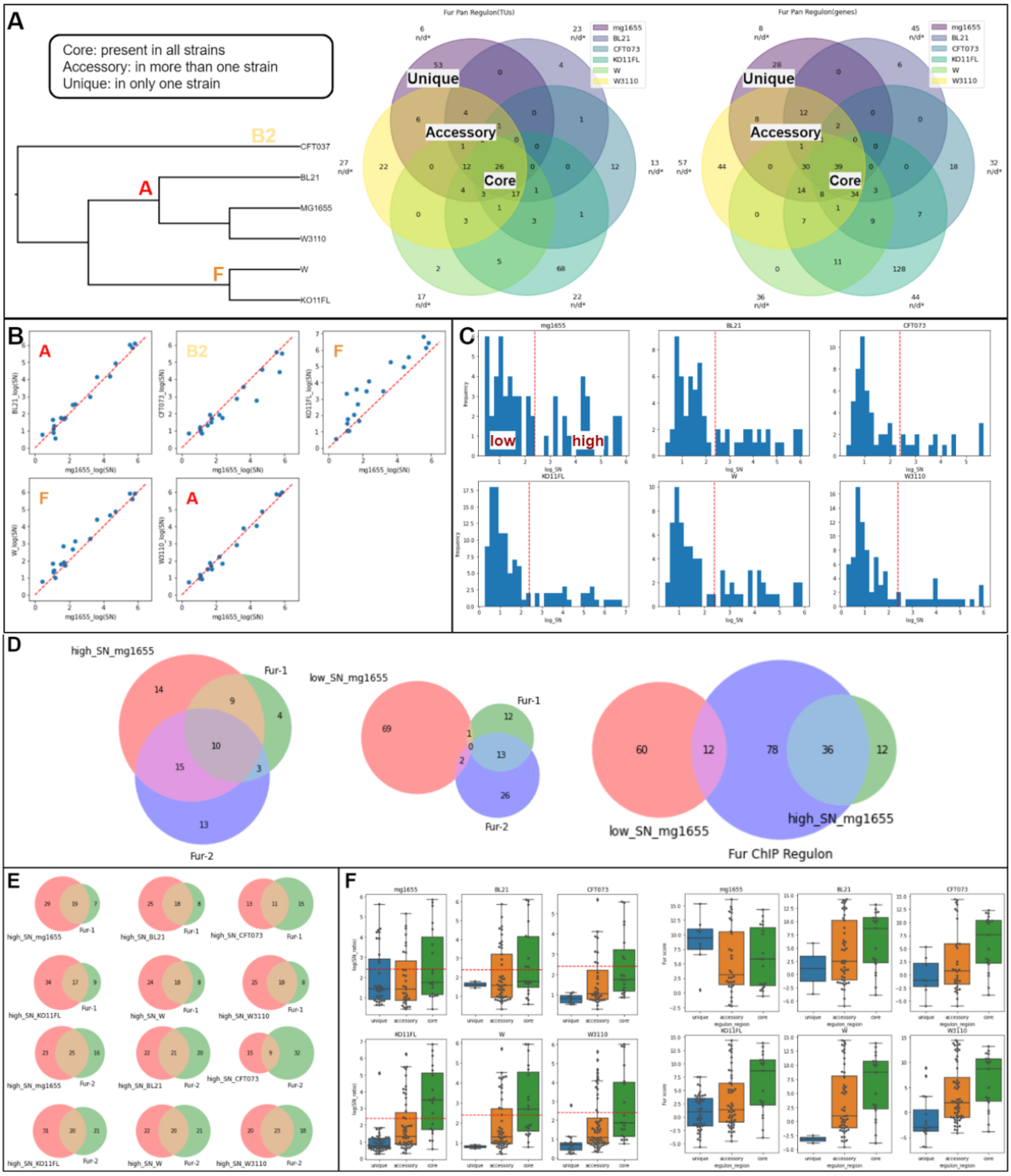

Figure 7. Fur pan-regulon analysis. (A) Fur pan regulons for both regulated genes and transcription units for 6 strains, with the phylogenetic tree displayed and unique, accessory, core regulons annotated. (B) Comparison of S/N ratios between MG1655 and all other 5 strains. (C) Histograms of S/N ratios in each strain. Bimodal distribution could be observed and a cutoff = 10 (2.3 in log-scale) used to separate high and low S/N ratios. (D) Venn diagrams showing relationships of genes with high/low S/N ratios and Fur ICA/ChIP TRNs for MG1655 strain. (E) Venn diagrams showing relationships of genes with high/low S/N ratios and Fur ICA TRNs for 6 strains. (F) Comparison of S/N ratios and Fur motif score: core > accessory > unique.

The classification result in Figure 8B demonstrates that features used in section 3.1 are sufficient to predict S/N ratios being high or low, and high correlation between motif scores and S/N ratios in Figure 8C suggests that motif scores are important indicators. Helix twist was discovered as another important indicator from feature importances of classification models (Figure 8D). The distributions of helix twist range at matched binding sites between genes of high and low S/N ratios show the tendency that gene promoters' binding sites with smaller Helix Twist range are more likely to have high S/N ratios (Figure 8E).

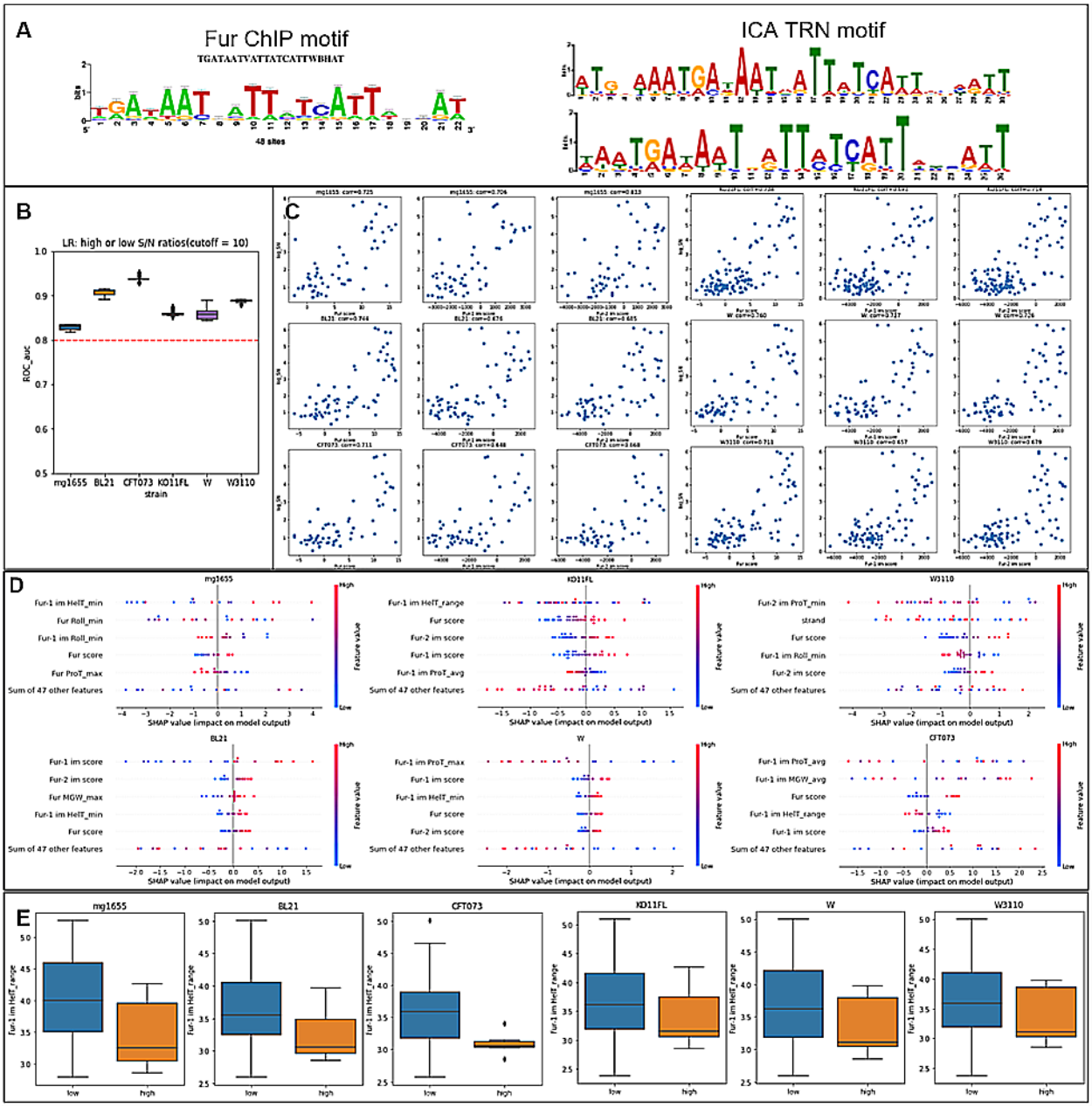

Figure 8. The Classification result of high/low S/N ratios for 6 strains. (A) Fur ChIP motif and 2 ICA regulon motifs in MG1655. (B) Assessment of classification models. (C) Correlations between S/N ratios and motif scores. (D) Feature importances: Motif scores and Helix Twist are two most important features to predict high or low S/N ratios. (E) Comparison of Helix Twist range at matched binding sites between genes of high and low S/N ratios. Smaller helix twist range indicates a higher S/N ratio.

In summary, the sequence basis of TRN's variation across strains lies in binding site strength, represented by motif scores. And the important features to determine S/N ratios, which is the regulation activity, are robust in multiple strains: motif scores and helix twist.

DISCUSSION

With respect to the current status of ICA TRN prediction, the classification workflow could predict over 50% ICA regulons' membership with high accuracy, but there are still a lot of potential improvements to be made. There were no motifs included in the data set of features for 30 ICA regulons, more data has to be curated. And promoter surveys should be carried out to find new discerning features. For example, RpoS ICA regulon is still hard to predict due to the fact that its motif is similar to that of RpoD, and thus new features of promoters bound by RpoS need to be discovered. Eventually, the ICA TRN prediction workflow will be able to reconstruct TRN with information extracted from the genome.

A better motif matching method than linear search with PSSM could be developed to improve the computation of motif scores and characterize binding site strength better. Though PSSM provides the probability of the segment being a binding site, the shape of the DNA sequence is not considered. There were already a few studies using DNA shape in binding site prediction, for example, improvement was observed in the machine learning model using both DNA sequence and shape to predict TF binding sites in ChIP-seq datasets [18]. Therefore, the prediction of TRN will become more accurate if the scoring function used in motif matching can be formulated with both PSSM and DNA shape included.

From the comparison between ICA and ChIP TRNs, it was found that ICA TRN is good at capturing useful information and supplementing the missing part of ChIP TRN. Fnr is not the only case that the regulon gets split, there are a few other cases, such as 5 RpoN associated ICA regulons and CysB -1,2,3 ICA regulons. Further investigation on cases where ICA/ChIP TRNs differ could provide insights into the relationship between the promoter sequence and resulting

transcription signal. If two methods to determine TRN are combined, a better representation of TRN could be constructed with more detailed knowledge.

In the analysis of Fur pan-regulon, promoters in other strains of E. coli must be predicted, but the method used in this study could only find promoters in highly conserved transcription units, and hence, many promoters' positions are unknown. The lack of promoters' data resulted in infeasibility of computing sequence features for certain genes, which led to filtration of measured S/N ratios. The final result was also affected, as in Figure 7F, many genes in MG1655's unique regulon of low S/N ratios were filtered due to missing transcription start sites. Therefore, a novel method to predict promoters is needed. If the scoring function using both PSSM and DNA shape mentioned above is applicable, it can predict -10/-35 boxes of sigma factors with high accuracy. Since the distance between those 2 boxes to the transcription start site is stable, the promoter is also determined in this case.

Overall, a reliable workflow to analyze transcription regulation activity in prokaryotes based on promoter sequence features was built, and this study provided a preliminary machine learning model to reconstruct the TRN with genomic information. Biophysical factors affecting the architecture of TRN and regulation activity were surveyed, offering directions for the data-driven genome design in synthetic biology to control cellular activity by tuning genomic objects such as promoters.

REFERENCE

[1] Browning, D. F., & Busby, S. J. (2004). The regulation of bacterial transcription initiation. Nature Reviews Microbiology, 2(1), 57-65. doi:10.1038/nrmicro787

[2] Martínez-Antonio, A., Janga, S. C., & Thieffry, D. (2008). Functional organisation of Escherichia coli transcriptional regulatory network. Journal of Molecular Biology, 381(1), 238-247. doi:10.1016/j.jmb.2008.05.0543.

[3] Myers, K. S., Park, D. M., Beauchene, N. A., & Kiley, P. J. (2015). Defining bacterial regulons using ChIP-seq. Methods, 86, 80-88. doi:10.1016/j.ymeth.2015.05.

[4] Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. Neural Networks, 13(4-5), 411-430. doi:10.1016/s0893-6080(00)00026-5

[5] Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K., Yang, L., King, Z., Palsson, B. O. (2019). The Escherichia coli transcriptome mostly consists of independently regulated modules. doi:10.1101/620799

[6] Gao, Y., Bang, I., Seif, Y., Nam, G., Sastry, A. V., Chen, K., Monk, J., Choudhary, K., Seo, S., Lee, E., Kim, D., Palsson, B. O. (2020). Reconstruction of Fur pan-regulon uncovers the complexity and diversity of transcriptional regulation in E. coli. doi:10.1101/2020.05.21.109694

[7] Lamoureux, C. R., Choudhary, K. S., King, Z. A., Sandberg, T. E., Gao, Y., Sastry, A. V., Phaneuf, P. V., Choe, D., Cho, B., Palsson, B. O. (2020). The BITOME: Digitized Genomic Features reveal fundamental genome organization. Nucleic Acids Research, 48(18), 10157-10163. doi:10.1093/nar/gkaa774

[8] Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., García-Sotelo, J., Alquicira-Hernández, K., Peña-Loredo,P., Collado-Vides, J. (2018). RegulonDB V 10.5: Tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. Nucleic Acids Research, 47(D1). doi:10.1093/nar/gky1077

[9] Wagner, R. (2000). The 'players' or cellular components necessary for transcription. In 1280753066 944877073 R. Wagner (Author), Transcription regulation in prokaryotes (pp. 16-57). Oxford: Oxford University Press.

[10] Warman, E. A., Singh, S. S., Gubieda, A. G., & Grainger, D. C. (2020). A non-canonical promoter element drives spurious transcription of horizontally acquired bacterial genes. Nucleic Acids Research, 48(9), 4891-4901. doi:10.1093/nar/gkaa244

[11] Lavrrar, J. L., & McIntosh, M. A. (2003). Architecture of a Fur Binding site: A comparative analysis. Journal of Bacteriology, 185(7), 2194-2202. doi:10.1128/jb.185.7.2194-2202.2003

[12] Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, "MEME SUITE: tools for motif discovery and searching", Nucleic Acids Research, 37:W202-W208, 2009.

[13] Park, D. M., Akhtar, M. S., Ansari, A. Z., Landick, R., & Kiley, P. J. (2013). The bacterial response regulator arca uses a diverse binding site architecture to regulate carbon oxidation globally. PLoS Genetics, 9(10). doi:10.1371/journal.pgen.1003839

[14] Chiu, T., Comoglio, F., Zhou, T., Yang, L., Paro, R., & Rohs, R. (2015). DNAshapeR: An R/Bioconductor package for DNA shape prediction and Feature encoding. Bioinformatics, 32(8), 1211-1213. doi:10.1093/bioinformatics/btv735

[15] Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S.,Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. Journal of Biomolecular Structure and Dynamics, 33(Sup1), 8-9. doi:10.1080/07391102.2015.1032554

[16] Scikit-learn: Machine Learning in Python, Pedregosa, F., Varoquaus, G., Gramfort, A., Michel, V., Thirion, B., JMLR 12, pp. 2825-2830, 2011.

[17] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

[18] Mathelier, A., Xin, B., Chiu, T., Yang, L., Rohs, R., & Wasserman, W. (2016). DNA shape features improve transcription factor binding site predictions in vivo. Cell Systems, 3(3). doi:10.1016/j.cels.2016.07.001