Open Access

OBM Genetics

Technical Note

Encoding, Regression, and Classification of Transcription Factors' Specificity and Methylation Effects

Zheng Zuo *

Fordyce lab, Department of Genetics, Stanford University, 450 Serra Mall, Stanford, CA 94305, US; E-Mail: zzuo@stanford.edu

* **Correspondence:** Zheng Zuo; Email: zzuo@stanford.edu

Academic Editor: Joep Geraedts

Special Issue: Advances in DNA Methylation

OBM Genetics
2021, volume 5, issue 3
doi:10.21926/obm.genet.2103134

Received: February 01, 2021
Accepted: August 04, 2021
Published: August 16, 2021

Abstract

The methylation effects on protein-DNA interactions, which can be perceived as a special kind of specificity of transcription factors, have been successfully quantified in the last years by various methods. In this work, I give a summary about the sequence encoding scheme, the underlying additive model about specificity and methylation sensitivity, and the regression strategy to analyze Methyl-Spec-seq data. Then I explain why given the current experimental setup, it is more appropriate to model the methylation effects based on pairwise comparison between individual unmethylated and methylated site, rather than the combined regression of all model parameters together. I also developed a computational package TFCookbook to demonstrate the analysis procedures step-by-step. At last, it is possible to classify the various types of methylation effects based on whether or not the consensus site contains CpG dinucleotide and whether methylation increase or decrease the binding affinity. Additionally, this specificity modeling and analysis strategy, can be extended to study other types of DNA modifications in general.

Keywords

Nucleotide encoding; methylation sensitivity; Spec-seq; Methyl-Spec-seq; TFCookbook; CTCF

1. Introduction

The DNA-binding specificity of some transcription factor (TF) refers to the information about its binding energy (ΔG) or affinity (K_D) to one sequence relative to all other possible DNA sequences. Given this information, we can predict where this TF binds *in vivo* within the genome, and more importantly, when some mutations or variations occur within the DNA binding domain of the TF or *cis*-regulatory elements (CREs), we can know how the underlying binding patterns change and what would be the functional consequences. Since there are different types of DNA modifications, e.g. methylation (mC), hydroxyl methylation (hmC), formylation (fC), carboxylation (caC), and adenylation etc., the specificity profile should not be restricted to four bases only. Cytosine methylation within CpG dinucleotide, as one of the most extensively studied modifications in vertebrates, drew lots of attention in the last decades for its diverse biological functions in epigenetic silencing [1, 2], imprinting control [3], and genome architecture maintenance [4, 5]. In the last years, multiple groups have developed high-throughput techniques to discover and quantify the methylation effects of many different TFs to their underlying binding sites, which seem to be a ubiquitous phenomenon among many human TFs [6, 7]. One of these methods, Methyl-Spec-seq [8], developed by the author of this work, can quantitatively assay methylation effects for a few thousand DNA sequences within one run and exhibits very good reproducibility with resolution down to ~0.2kT. The original Methyl-Spec-seq paper described the experimental procedures and the results for a few TFs, but did not present the data analysis protocol to model the methylation effects in detail. In this work, I give a summary about the sequence encoding scheme, specificity and methylation model, and regression strategy. Using CTCF as case example, I show that it is more realistic to model the methylation effects based on pairwise comparison between individual methylated and unmethylated site, instead of the combined regression of all sequences together, which would produce biased results even in the absence of methylation effect at all. To implement such two-step modelling strategy, I developed a computational package TFCookbook to demonstrate the protocol step-by-step for CTCF and HoxB13. At last, it is feasible to classify the various types of methylation effects based on the underlying consensus site and the direction of affinity change. Overall, our experimental method and analysis strategy can be extended to study other types of DNA modifications in general.

2. Results

2.1 The Canonical 3L+1 Coding Scheme is Used to Model the Specificity of Protein-DNA Interactions

Often people present specificity profile of a TF as position weight matrix (PWM) [9] and visualize it as some logo, either in terms of information content or log-odds ratios, which are correlated with but not directly proportional to the binding energy or affinity of each DNA sequence. As high-throughput biophysical measurements become increasingly available, it is advantageous to directly model and present the specificity in terms of Gibbs-free energy (ΔG) with kT or kcal/mol as units, which we call position energy matrix (PEM). All such models natively assume position-independence and additivity, i.e., the binding energy of each sequence can be estimated by the sum of

contributions of individual bases at each position, thus for any sequence of length L, we can build some parametric PEM with 4L independent parameters ($\beta_{1C}, \beta_{1G}, \beta_{1A}, \beta_{1T}, \beta_{2C}, \beta_{2G}$, etc.) to estimate its binding energy, which we call 4L model. For example, for sequence TACG, we can estimate its energy as

4L model:

$$\Delta\hat{G}_{TACG} = \beta_{1T} + \beta_{2A} + \beta_{3C} + \beta_{4G} \tag{1}$$

Here the β_{1T} is interpreted as the energy contribution of base T at position 1, and so on. However, the drawback is that there exists more than one equivalent models to predict the exact same energy for every site [10], e.g., we can subtract a fixed number δ from each parameter at position 1 and add it back to each parameter at position 3, and now the new model gives exactly the same predictions as the original one.

Equivalent 4L model:

$$\begin{aligned}\Delta\hat{G}_{TACG} &= (\beta_{1T} - \delta) + \beta_{2A} + (\beta_{3C} + \delta) + \beta_{4G} \\ &= \beta'_{1T} + \beta_{2A} + \beta'_{3C} + \beta_{4G}\end{aligned} \tag{2}$$

To fix such redundancy and make each model unique, we can pick some site as reference, introduce an intercept parameter ε as the energy level for this reference site, and have parameters that quantify the energy contribution of other bases relative to the corresponding base within the reference site, e.g., if we choose CCCC as the reference site, then the energy of TACG can be predicted as:

(3L+1) model:

$$\Delta\hat{G}_{TACG} = \varepsilon_{CCCC} + \beta_{1CT} + \beta_{2CA} + \beta_{4CG} \tag{3}$$

Here ε_{CCCC} is the estimated energy for reference site CCCC, whereas β_{1CT} can be interpreted as the energy cost to switch base C to T at position 1, and so on. Note that $\beta_{1CC}, \beta_{2CC}, \beta_{3CC}, \beta_{4CC}$ are all zeros, thus each position x has three independent parameters ($\beta_{xCG}, \beta_{xCA}, \beta_{xCT}$), so there are totally 3L+1 parameters for any model for sequences of fixed length L, which is why we call it 3L+1 model. Under this model, we can encode any sequence of length L into some unique binary vector of length 3L, and predict its energy as the inner product of this binary coefficient vector $\vec{\mathrm{X}}$ with corresponding PEM model parameters plus the reference site energy, e.g.

$$\begin{aligned}\Delta\hat{G}_{TACG} &= \Delta\hat{G}_{001\,010\,000\,100} = \varepsilon_{CCCC} + \vec{X}\cdot\vec{\beta} \\ &= \varepsilon_{CCCC} + \\ &0\cdot\beta_{1CG} + 0\cdot\beta_{1CA} + 1\cdot\beta_{1CT} + 0\cdot\beta_{2CG} + 1\cdot\beta_{2CA} + 0\cdot\beta_{2CT} + \\ &0\cdot\beta_{3CG} + 0\cdot\beta_{3CA} + 0\cdot\beta_{3CT} + 1\cdot\beta_{4CG} + 0\cdot\beta_{4CA} + 0\cdot\beta_{4CT} \\ &= \varepsilon_{CCCC} + \beta_{1CT} + \beta_{2CA} + \beta_{4CG}\end{aligned} \tag{4}$$

If we perform some high-throughput biophysical experiment, e.g., PBM, HT-SELEX, Spec-seq, and MITOMI, and get the relative binding affinity or energy values ΔG_{S_i} for a collection of sequences S_i, we can define the residual sum of squares (R.S.S.) as in Equation 5 to characterize the total deviation between observed and predicted energy values by any particular PEM model.

$$\text{Residual Sum of Squares (R.S.S.)} = \sum_{S_i} \left(\underset{\text{Observed Energy}}{\Delta G_{S_i}} - \underset{\text{Predicted Energy}}{\Delta \hat{G}_{S_i}} \right)^2 = \sum_{S_i} \left(\Delta G_{S_i} - \varepsilon_{CCCC} - \sum_{j=1}^{3L} x_{ij}\beta_j \right)^2 \quad (5)$$

where x_{ij} is the binary coding coefficient of sequence S_i onto parameter β_j

Under this 3L+1 coding scheme, it becomes feasible to use least-square-fit to find a unique, optimal model or one set of parameters such that the R.S.S gets minimized, which is usually visualized in energy logo format (To make the reference base C visible in the logo, we usually shift PEM parameters within the same position by a fixed amount so that all bases add up to zero). Intuitively, we can interpret the parameters of this optimal PEM as the best estimate of contribution of each base at corresponding position. It is easy to show that the prediction by this optimal PEM does not depend on the choice of reference site, and the reference site does not even need to be included in the measurement series. This 3L+1 encoding and regression strategy is the default method for modelling the specificity in our regular Spec-seq work, and the software package TFCookbook is developed on this premise.

2.2 (3+2)L+1 Coding Scheme Can be Used to Model CpG Methylation Effects of Transcription Factors

When some cytosine base gets methylated, we can represent it either as M or W, depending on whether it is in the upper or lower strand of the DNA duplex. Given a TF of interest, the methylation effects for some CpG dinucleotide within its binding site can be defined as the binding energy difference between the methylated and unmethylated sites. The most intuitive way to model methylation effects is to introduce two extra parameters β_{xCM} and β_{xGW} at each position x, and interpret them as the extra amount of energy cost when the upper base C gets converted to methylated C (or M), and the lower base C (upper G) gets converted to methylated C (or W) respectively, as in Equation 6. With 2L extra parameters ($\beta_{1CM}, \beta_{1GW}, \beta_{2CM}, \beta_{2GW}$, etc.), we name the new model as (3+2)L+1 model. In some Methyl-Spec-seq experiment like CTCF, we used methyltransferase M.SssI alone to methylate CpG dinucleotide, so CpG dinucleotides are always methylated together to MW. There is no way to distinguish strand-specific methylation effects, so we can simplify above model using only one extra parameter β_{xMW} at position x, representing the energy cost to substitute CG to MW at position (x, x+1), as in Equation 7. For such (3+2)L+1 and (3+1)L+1 models, we can extend the binary coding vector correspondingly as in Table 1, and perform multiple linear regression over measurement results the same way as 3L+1 case to fit the data.

(3+2)L+1 model:

$$\begin{aligned}\Delta\widehat{\mathrm{G}}_{\mathrm{TAMW}} &= \Delta\widehat{\mathrm{G}}_{00100\ 01000\ 00010\ 10001}\\ &=\varepsilon_{\mathrm{CCCC}}+\\ &0\cdot\beta_{1\mathrm{CG}}+0\cdot\beta_{1\mathrm{CA}}+1\cdot\beta_{1\mathrm{CT}}+0\cdot\beta_{1\mathrm{CM}}+0\cdot\beta_{1\mathrm{GW}}+\\ &0\cdot\beta_{2\mathrm{CG}}+1\cdot\beta_{2\mathrm{CA}}+0\cdot\beta_{2\mathrm{CT}}+0\cdot\beta_{2\mathrm{CM}}+0\cdot\beta_{2\mathrm{GW}}+\\ &0\cdot\beta_{3\mathrm{CG}}+0\cdot\beta_{3\mathrm{CA}}+0\cdot\beta_{3\mathrm{CT}}+1\cdot\beta_{3\mathrm{CM}}+0\cdot\beta_{3\mathrm{GW}}+\\ &1\cdot\beta_{4\mathrm{CG}}+0\cdot\beta_{4\mathrm{CA}}+0\cdot\beta_{4\mathrm{CT}}+0\cdot\beta_{4\mathrm{CM}}+1\cdot\beta_{4\mathrm{GW}}\\ &=\varepsilon_{\mathrm{CCCC}}+\beta_{1\mathrm{CT}}+\beta_{2\mathrm{CA}}+\beta_{3\mathrm{CM}}+\beta_{4\mathrm{CG}}+\beta_{4\mathrm{GW}}\\ &=\Delta\widehat{\mathrm{G}}_{\mathrm{TACG}}+\beta_{3\mathrm{CM}}+\beta_{4\mathrm{GW}}\end{aligned} \tag{6}$$

(3+1)L+1 model:

$$\begin{aligned}\Delta\hat{G}_{TAMW} &= \Delta\hat{G}_{0010\ 0100\ 0001\ 1000}\\ &=\varepsilon_{CCCC}+\beta_{1CT}+\beta_{2CA}+\beta_{3MW}+\beta_{4CG}\\ &=\Delta\hat{G}_{TACG}+\beta_{3MW}\end{aligned} \tag{7}$$

Table 1 Binary coefficient vectors for some representative sequences under different coding schemes.

Sequence	4L encoding	3L+1 encoding	(3+2)L+1 encoding	(3+1)L+1 encoding
TACG	0001 0010 1000 0100	001 010 000 100	00100 01000 00000 10000	0010 0100 0000 1000
AACG	0010 0010 1000 0100	010 010 000 100	01000 01000 00000 10000	0100 0100 0000 1000
TAMW	N/A	N/A	00100 01000 00010 10001	0010 0100 0001 1000

2.3 Regression of CTCF Methyl-Spec-Seq Data with 4L+1 Parameters All in Once Produces Inaccurate Estimates about Methylation Effects of CTCF

CTCF is known to be sensitive to mCpG within its core binding site [5], which was reported to be important for imprinting maintenance [4]. In our previous Methyl-Spec-seq experiment for CTCF(F1-F9), we designed tandem random dsDNA libraries to cover the core CTCF binding site recognized by fingers 3 to 7 of mouse CTCF, as in Figure 1A. In the meantime, we included M.SssI treated DNA libraries with an alternative barcode at flanking position 19, therefore for each sequence S_i in the untreated libraries, there is one corresponding sequence M_i in the M.SssI treated libraries, where CpG gets methylated to MpW. A simple comparison between all such pairs (Figure 1B) reveals that among all tested low binding energy or high affinity sites, only site C**CG**GTAGGGGGCACTA shows significantly increased binding energy up to 1kT when its CpG at position 2 gets methylated, whereas other positions show small effects that cannot be definitively determined by current measurement resolution of Spec-seq around 0.2kT, thus we expect a realistic PEM model should have some non-

zero, positive β_{2MW} parameter, and all other methyl parameters are negligible. Nonetheless, when we perform multiple linear regression over 4L+1 parameters altogether to find a set of parameters minimizing the deviation between predicted and observed results, we end up with non-zero methyl parameters at many different positions, and most surprisingly, β_{2MW} are significantly smaller than what we expect (Figure 1D).

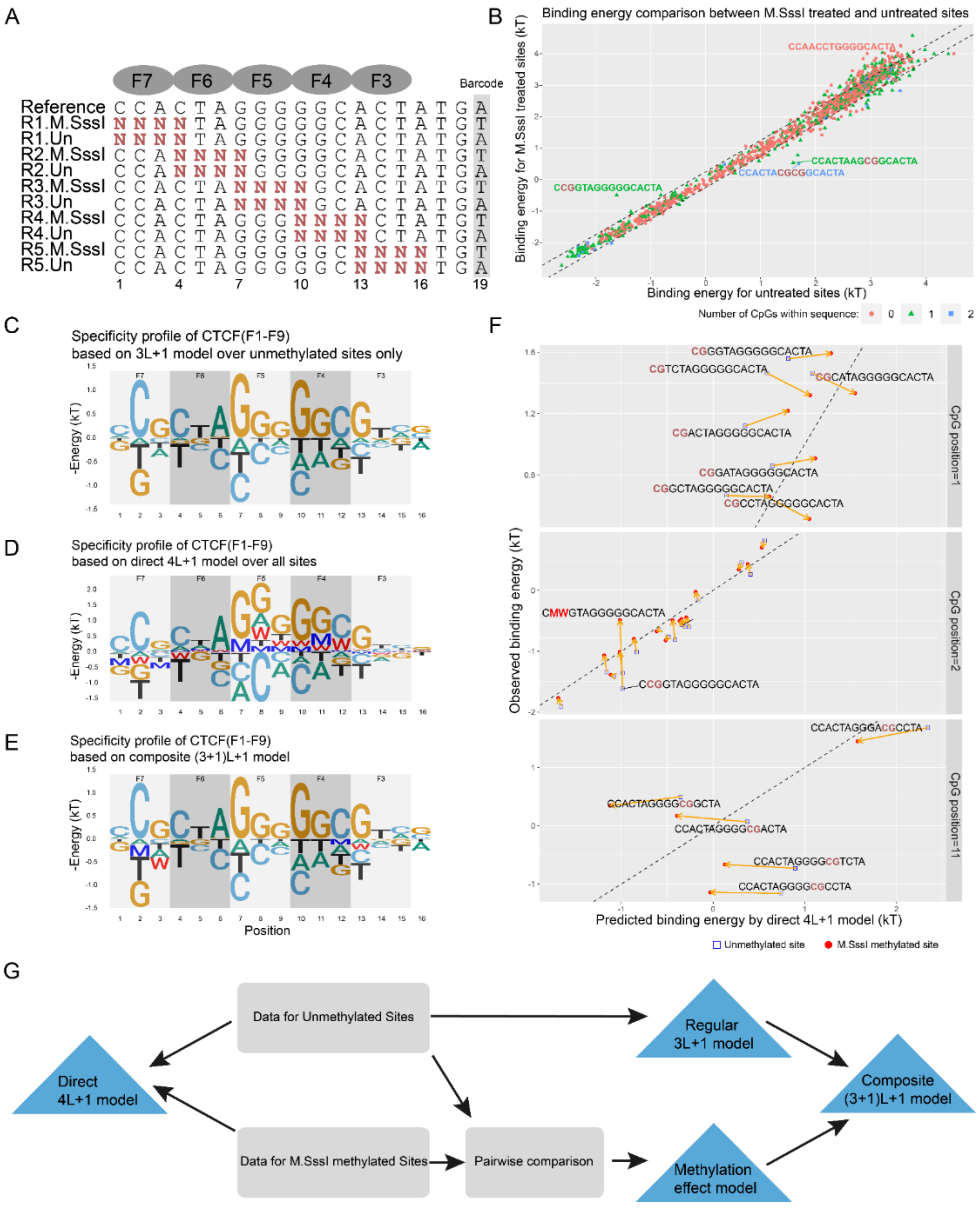

Figure 1 A) Methyl-Spec-seq libraries design. R1-R5.Un are not treated by M.SssI methyltransferase, whereas R1-R5.M.SssI are treated by M.SssI and those CpG-containing sites get methylated. **B)** Comparison of observed binding energy values between M.SssI treated and untreated sequences. Dashed lines are 0.25kT deviation bounds. **C-E)** Energy logos produced by regular 3L+1 model, all-in-one 4L+1 model, and composite (3+1)L+1 model respectively. **F)** Comparison of observed binding energy with predicted values by 4L+1 model for methylated/unmethylated pairs of sequences at different CpG positions (1, 2, and 11). **G)** Overview of the workflow to build different kinds of specificity and methylation effects models.

To figure out what causes this discrepancy, we can draw diagrams showing the predicted versus observed binding energy values for CpG-containing sites at different position, as in Figure 1F. We think that two reasons explain this discrepancy. First, among all sites containing CpG at position 2, only sequence C**CG**GTAGGGGGCACTA shows a strong methylation effect up to 1kT, whereas other ones at the same position give relatively insignificant results, possibly due to various reasons, e.g., context-dependent methylation effects or incomplete methyltransferase efficiency. So overall, this significant, single site methylation effect is 'buried' inside other insignificant ones and thus did not show up in the direct 4L+1 PEM model. Second, the all-in-one regression method aims to find a set of 4L+1 parameters so that the total R.S.S. gets minimized, which doesn't aim to fit the methylation effects alone. So very often, the model uses non-zero methylation parameters to fit any unexplained portion between the observed and predicted values, even though those unexplained portions are just intrinsic limitations of position-independent, additive models and there is no methylation effect involved, which are the cases for CTCF at position 1 and 11 (Figure 1F). To illustrate this point more intuitively, we can even create some artificial data set for CTCF by setting every M.SssI methylated site exactly to the same value as the untreated one. Under this circumstance, there should be no methylation effect at all, but when we perform all-in-one regression over this artificial dataset, we still get non-zero methyl parameters at many different positions, highlighting the intrinsic limitation of this all-in-one modelling strategy (Supplemental S1).

2.4 Separate Regression of Methylation-Specific Parameters over Pairwise Compared Data Gives Good Estimate about Methylation Sensitivity of CTCF

Since methylation sensitivity is a special kind of specificity and have important biological functions, it is worth performing separate regression to better estimate it at individual position. We can construct a methyl-specific R.S.S. (Equation 8) to quantify the total deviation between the observed and predicted methyl effects by any particular (3+1)L+1 PEM model. Note that this $\text{R.S.S.}_{\text{Methyl}}$ operates on a set of pairwise compared data between corresponding methylated and unmethylated sites, e.g., between CCGGATC and CMWGATC, which matches our experiment design of Methyl-Spec-seq well. For each pair of sequences (M_i, S_i), all parameters unrelated to methyl effects cancel out against each other so overall this metric only involves methyl parameters.

$$\begin{aligned}\text{R.S.S.}_{\text{Methyl}} &= \sum_{(M_i,S_i)} \Big[\underbrace{(\Delta G_{M_i} - \Delta G_{S_i})}_{\text{Observed Methyl effect}} - \underbrace{(\Delta \hat{G}_{M_i} - \Delta \hat{G}_{S_i})}_{\text{Predicted Methyl effect}} \Big]^2 \\ &= \sum_{(M_i,S_i)} \left(\Delta G_{M_i} - \Delta G_{S_i} - \sum_{j=1}^{L} x_{ij}\beta_{jMW} \right)^2 \end{aligned} \tag{8}$$

where M_i is the M.SssI methylated site of S_i, and x_{ij} is the binary coding coefficient depending on whether sequence S_i has CpG dinucleotide at position (j, j+1).

It is easy to perform regression of non-methyl parameters over unmethylated sites only to derive a regular 3L+1 model to characterize specificity. Regression of methyl parameters over pairwise compared data allows us to derive a methylation effects model with L extra parameters separately.

By combining these two models we are able to construct a composite (3+1)L+1 model for CTCF and visualize it as in Figure 1E. Since our measurement resolution is around 0.25kT, it is legitimate to set a cut-off value at 0.25kT and drop those methyl effects below this threshold, so only those significant methyl effects show up in our final result. The end result matches our expectation quite well and clearly shows that CTCF is negatively regulated by methylation of CpG dinucleotide at position 2, which is also consistent with existing literatures. In Supplemental Material, we use (3+2)L+1 model and the same strategy to analyze the specificity and methylation effects of HoxB13, which was tested by both strand-specific methylation and M.SssI-mediated methylation, and found that HoxB13 contains two independent motifs (or mode of recognition) and exhibits strand-specific methylation effect only on the upper strand.

2.5 Classification of Different Types of Methylation Sensitivity of Transcription Factors

Since there have been numerous established examples of methylation-dependent TF binding, we can classify them based on two criteria---Does methylated CpG increase or decrease binding affinity? Is CpG dinucleotide the optimal binding site for the unmethylated sequence? Clearly for TFs like Kaiso [11], YY1 [12], and Gli1 (internal position 13, 14) [8], despite CpG is not the strongest unmethylated binding site at relevant positions, methylation still can modulate binding affinity when those positions have CpG as suboptimal sequence. Thus, our study for methylation effects of TFs should never be restricted to those TFs with CpG within their logos and a lot more needs to be explored.

Table 2 Types of Methylation effects.

	Is CpG the optimal binding site?	
	Yes	No
Increase affinity upon methylation	Type I (ZFP57, HoxB13)	Type II (Kaiso)
Decrease affinity upon methylation	Type III (CTCF, AP1)	Type IV (YY1, Gli1)

3. Discussion

As George Box said, “All models are wrong, but some are useful.” Ideally, we want to have one simple model to explain everything we observed, but in reality, a ‘good’ model involves rational experimental design, proper knowledge about the measurement resolution, and most importantly, clear understanding about what we want to learn. In this work, we showed that all-in-one linear regression does not give the proper answer about the methyl effects of CTCF, primarily because the model ‘squeezes’ some unexplained portion of data onto the methylation dimension, which ends up in a sub-optimal result for methylation effects, therefore separate modeling of methylation effects is required. There are other scenarios that linear, additive model fails for alternative reasons. For example in the HoxB13 case, we and Taipale group [6] notice that single PWM is insufficient to explain the observed data, particularly for small portions of sequences, thus there must be alternative structural conformation or recognition mode to explain that unexplained portions of

sequences, which is proved by structural studies later on [13]. Actually, this is not unique case and similar observation are made in other TFs, like FoxN3 [14], thus this insufficiency of model helps us learn more about how TFs work.

Overall, we want to use this work to show the importance of clear understanding about the limitation of one type of modeling and careful interpretation about experimental results. Similar encoding and analysis strategy should be able to be extended to study of other types of DNA modifications [15] without much difficulty.

Additional Materials

The following additional materials are uploaded at the page of this paper.

1. Supplemental S1: Step-by-step procedures for the analysis of CTCF and HoxB13. The source code of TFCookbook package can be accessed through GitHub at https://github.com/zeropin/TFCookbook.

Acknowledgement

The author really wants to thank anonymous reviewers for their helpful feedback to revise this work.

Author Contributions

Z.Z. developed TFCookbook package, analyzed previously published data, and wrote the manuscript.

Competing Interests

The author has declared that no competing interests exist.

Reference

1. Razin A. CpG methylation, chromatin structure and gene silencing – a three-way connection. EMBO J. 1998; 17: 4905-4908.
2. Robert MF, Morin S, Beaulieu N, Gauthier F, Chute IC, Barsalou A, et al. DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. Nat Genet. 2003; 33: 61-65.
3. Quenneville S, Verde G, Corsinotti A, Kapopoulou A, Jakobsson J, Offner S, et al. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. Mol Cell. 2011; 44: 361-372.
4. Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature. 2000; 405: 486-489.
5. Hashimoto H, Wang D, Horton JR, Zhang X, Corces VG, Cheng X. Structural basis for the versatile and methylation-dependent binding of CTCF to DNA. Mol Cell. 2017; 66: 711-720.

6. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science. 2017; 356. doi: 10.1126/science.aaj2239.
7. Kribelbauer JF, Laptenko O, Chen S, Martini GD, Freed-Pastor WA, Prives C, et al. Quantitative analysis of the DNA methylation sensitivity of transcription factor complexes. Cell Rep. 2017; 19: 2383-2395.
8. Zuo Z, Roy B, Chang YK, Granas D, Stormo GD. Measuring quantitative effects of methylation on transcription factor–DNA binding affinity. Sci Adv. 2017; 3: eaao1799.
9. Stormo GD. DNA binding sites: Representation and discovery. Bioinformatics. 2000; 16: 16-23.
10. Stormo GD. Maximally efficient modeling of DNA sequence motifs at all levels of complexity. Genetics. 2011; 187: 1219-1224.
11. Prokhortchouk A, Hendrich B, Jørgensen H, Ruzov A, Wilm M, Georgiev G, et al. The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. Genes Dev. 2001; 15: 1613-1618.
12. Kim J, Kollhoff A, Bergmann A, Stubbs L. Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3. Hum Mol Genet. 2003; 12: 233-245.
13. Morgunova E, Yin Y, Das PK, Jolma A, Zhu F, Popov A, et al. Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. Elife. 2018; 7: e32963.
14. Rogers JM, Waters CT, Seegar TC, Jarrett SM, Hallworth AN, Blacklow SC, et al. Bispecific forkhead transcription factor FoxN3 recognizes two distinct motifs with different DNA shapes. Mol Cell. 2019; 74: 245-253.
15. Xiong J, Zhang Z, Chen J, Huang H, Xu Y, Ding X, et al. Cooperative action between SALL4A and TET proteins in stepwise oxidation of 5-methylcytosine. Mol Cell. 2016; 64: 913-925.