9392–9403 *Nucleic Acids Research, 2012, Vol. 40, No. 19* *Published online 16 August 2012*
doi:10.1093/nar/gks749

Subtypes of associated protein–DNA (Transcription Factor-Transcription Factor Binding Site) patterns

Tak-Ming Chan[1,*], Kwong-Sak Leung[1], Kin-Hong Lee[1], Man-Hon Wong[1], Terrence Chi-Kong Lau[2,*] and Stephen Kwok-Wing Tsui[3,4]

[1]Department of Computer Science & Engineering, The Chinese University of Hong Kong, Shatin, N. T., [2]Department of Biology and Chemistry, The City University of Hong Kong, Kowloon, [3]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, N. T. and [4]Hong Kong Bioinformatics Centre, Shatin, N. T., Hong Kong

Received January 3, 2012; Revised July 6, 2012; Accepted July 16, 2012

ABSTRACT

In protein–DNA interactions, particularly transcription factor (TF) and transcription factor binding site (TFBS) bindings, associated residue variations form patterns denoted as subtypes. Subtypes may lead to changed binding preferences, distinguish conserved from flexible binding residues and reveal novel binding mechanisms. However, subtypes must be studied in the context of core bindings. While solving 3D structures would require huge experimental efforts, recent sequence-based associated TF-TFBS pattern discovery has shown to be promising, upon which a large-scale subtype study is possible and desirable. In this article, we investigate residue-varying subtypes based on associated TF-TFBS patterns. By re-categorizing the patterns with respect to varying TF amino acids, statistically significant (*P* values ≤ 0.005) subtypes leading to varying TFBS patterns are discovered without using TF family or domain annotations. Resultant subtypes have various biological meanings. The subtypes reflect familial and functional properties and exhibit changed binding preferences supported by 3D structures. Conserved residues critical for maintaining TF-TFBS bindings are revealed by analyzing the subtypes. In-depth analysis on the subtype pair PKVVIL-CACGTG versus PKVEIL-CAGCTG shows the V/E variation is indicative for distinguishing Myc from MRF families. Discovered from sequences only, the TF-TFBS subtypes are informative and promising for more biological findings, complementing and extending recent one-sided subtype and familial studies with comprehensive evidence.

INTRODUCTION

Protein–DNA interactions play a central role in genetic activities (1,2). In particular, transcription factor (TF, the protein side) and transcription factor binding site (TFBS, the DNA side) bindings are critical and primary protein–DNA interactions to be deciphered for gene regulation. So TF-TFBS bindings will be our focus throughout the article. Despite the great variations shown among different whole-length TF and TFBS sequences, part of them are conserved as TF binding domains (tens to hundreds of residues) and TFBS motifs (usually several to 20 residues), respectively. Within the distance of forming hydrogen bonds, short TF and TFBS subsequences show more conserved patterns. These associated short binding subsequences (within 10 residues; 6–8 in our experiments) of both TFs and TFBSs surrounding the interacting bonds are denoted as binding cores. However, predicting these short binding cores on both the TF and TFBS sides from sequences only is very challenging.

Amino acid residue variations in the TF binding cores may lead to intriguing different corresponding TFBS sub-patterns. For example, [A/P]KV[E/V]IL-CA[C/G][C/G]TG may be found to be TF-TFBS binding cores. Specifically, PKVEIL may bind to CAG[C/G]TG, whereas PKVVIL to CACGTG. We denote such associated TF-TFBS residue variations as subtypes, which should be studied in the context of associated TF-TFBS core bindings. 'PKVEIL-CAG[C/G]TG versus PKVVIL-CACGTG (3rd column)' and 'PKVEIL-CAG[C/G]TG versus PKVVIL-CACGTG (4th column)' are two related (column-specific) subtype pairs. Subtypes can reflect familial specificities, exhibit changed binding preferences, distinguish conserved residues from flexible ones and reveal novel binding mechanisms. Although such high-resolution details are usually extracted from 3D structures with huge experimental efforts, abundant low-resolution binding sequence data can be exploited to

*To whom correspondence should be addressed. Tel: +852 39434259; Fax: +852 39435024; Email: tmchan@cse.cuhk.edu.hk
Correspondence may also be addressed to Terrence Chi-Kong Lau. Tel: +852 34429327; Fax: +852 34420522; Email: chiklau@cityu.edu.hk

predict both testable TF-TFBS binding cores and subtypes.

In this article, we for the first time introduce and study residue varying subtypes based on our sequence-based associated TF-TFBS pattern discovery (3). The brief review is first given in the following sub-sections. TF-TFBS subtype discovery methods are detailed in 'Materials and Methods' section. Experimental results and verifications are reported in 'Results and Analysis' section, before the final 'Discussion and Conclusion' section.

TF-TFBS bindings in gene regulation

Because of functional importance of TF-TFBS bindings on regulation, core interaction subsequences from bound TFs and TFBSs are less likely to mutate and exhibit recognizable patterns (i.e. being conserved) from similar TF-TFBS bindings. The short conserved subsequence patterns on either TF or TFBS side are called motifs. TFs have relatively long conserved regions called domains of up to hundreds of amino acids (AA), but the core interaction subsequences interacting with TFBSs are shown to be highly specific (3,4). TFBS motifs are usually short [within 10 base pairs (bp)], and long motifs (up to 20 bp) are usually composites of short patterns separated by non-conserved gaps.

Existing data

Experiments to determine TF-TFBS bindings at the sequence level include the traditional DNA footprinting, gel electrophoresis and the new chromatin immunoprecipitation (ChIP) followed by chip (-chip) or sequencing (-seq) technology (5,6). For the resultant TF-TFBS binding sequence data, the resolution for a whole TF is hundreds of AA without knowing the binding domains, and the resolution for TFBSs is tens to hundreds of bp depending on experiment techniques. Although sequence level data contain noises and do not describe core protein–DNA interactions directly, they serve as the most widely available information for discovering elaborate binding patterns (motifs) based on sequence conservation.

TRANSFAC (7) is one of the largest and most representative databases for sequence-level binding data in regulation, including TFs, TFBSs and nucleotide distribution matrices of the TFBSs (TFBS motifs). The data are annotated and curated from peer-reviewed and experimentally proved publications. ChIP-Seq technology provides high-throughput and precise TF-TFBS binding sequence data *in vivo* for discovering TFBS motifs (8). High-quality ChIP-Seq motifs can serve as independent and indirect verification for our subtypes.

It is much more expensive and laborious to extract high-resolution 3D protein–DNA interaction (TF-TFBS binding) structures with X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopic analysis. The experiments provide binding data at atom level and clearly show interacting residues on both the protein (TF) and DNA (TFBS) sides. The Protein Data Bank (PDB) (9) is the most representative repository for such data. However, the available 3D structures are very limited compared with sequence data. On the other hand, PDB data serve as valuable verification sources for putative associated interaction patterns.

Existing methods to study protein–DNA interactions

On 3D structure data, 'one-to-one binding codes' between single amino acids and nucleotides (1,10) for protein–DNA interactions have been sought, followed by 'training-based methods' predicting protein binding residues (11,12). These studies are mainly constrained by the limited amount of 3D structures and only predict bindings of individual residues.

On sequence data, the early computational attempt has been 'motif discovery' of either TFs or TFBSs, with the inspiring subtype discovery (13) focusing on the subtle variations within single TFBS motifs. Recent TF-TFBS 'associated pattern discovery' from large-scale binding sequence data demonstrates a very promising direction, with 3D structure verification and novel predictions. A few novel studies on 'TF-TFBS binding co-evolution/variation' have also been proposed. They are briefly reviewed as follows:

Motif discovery

By exploiting the similar subsequences (i.e. conserved patterns) of TF/TFBS, motif discovery has long been studied with certain success. Motifs are usually represented as consensus strings or position weight matrices (PWMs) of the amino acid/nucleotide distributions (14). Recently, TFBS motif discovery has been extended to subtype discovery as groups of nucleotide variants (subtypes) contribute to distinct modes of regulation (13). Besides the existing challenges (15,16), a significant limitation of TF or TFBS motif discovery is the lack of linkage between to directly reveal the binding counterpart (TF-TFBS) relationship. On the other hand, this article considers both TFs and TFBSs beyond one-sided motif discovery and investigates binding subtypes on large-scale experiment-verified binding data (i.e. TRANSFAC) to provide insights into motifs and detailed binding mechanisms. Various TFBS motif (PWM) comparison methods have been developed (17–19), some of which are readily employed (e.g. Euclidean distance and Pearson chi-square test) while others can be explored in our future study.

Associated pattern discovery

Being the most widely available data, TF-TFBS binding sequences are better exploited on both TF and TFBS sides, than on only one side, for associated patterns to reveal intriguing binding mechanisms (20). Recent association rule mining (4) from TRANSFAC discovers exact TF-TFBS patterns verified on both literature and PDB 3D structures. More recently, approximate associated TF-TFBS pattern discovery (3) employing the probabilistic model significantly expands the verifiable patterns and outperforms traditional motif discovery methods, if they are recruited in the TF part of the task. Associated TF-TFBS patterns provide more general information than individual TF-TFBS binding records. As multiple binding records are included in one associated patterns, we

alleviate the limitation imposed by insufficient data in family-based and individual record-based studies.

Binding co-evolution/variation

There are also two novel studies addressing correlated evolution/co-variations of TF-TFBS bindings from sequences (21,22). Positive results are reported, and they are related to our proposed work. Our TF-TFBS binding subtype discovery distinguishes itself from the studies from several aspects:

- The previous studies focus on a small number (3–4) of specific TF families. Our study is general and large scale on the whole TRANSFAC database (across all possible families) and produces more biologically interesting case studies and novel results.
- More importantly, although the previous studies rely on known TF domain information, which may be limited [mainly around 10–20 samples per dataset in (22)], we work on computationally discovered approximate associated TF-TFBS patterns without requiring such annotations. The associated patterns are more conserved, facilitating more convenient analysis than degenerate aligned patterns.
- The previous major results are about statistical significances (21,22), and case studies are established upon literature support only (21). We show not only statistically significances but also evaluations extensively on PDB 3D structures. The previous case studies (21,22) have been fully covered and more interesting and novel examples are presented in this study (summary in Supplementary Data).

Although the previous studies are not directly comparable with our work here, our study complements and enriches them with much more comprehensive evidence and in-depth 3D analysis.

MATERIALS AND METHODS

In this section, we first present the TRANSFAC data, introduce the associated TF-TFBS pattern discovery, elaborate the subtype discovery methodology and then describe the analysis and verification procedures. The overview is shown in Figure 1. The corresponding results to the methods are given in the next 'Results and Analysis' section.

TRANSFAC data and associated TF-TFBS pattern discovery

Following our previous work (3), we employ TRANSFAC Professional ver 2009.4 (7), which contains 13 682 TF entries (7664 with protein sequences) and 1225 matrices of the TFBS nucleotide distributions, i.e. TFBS motif consensuses. Each TF is associated with the set of TFBSs it binds to and the corresponding TFBS consensus. With the consensus dissimilarity threshold

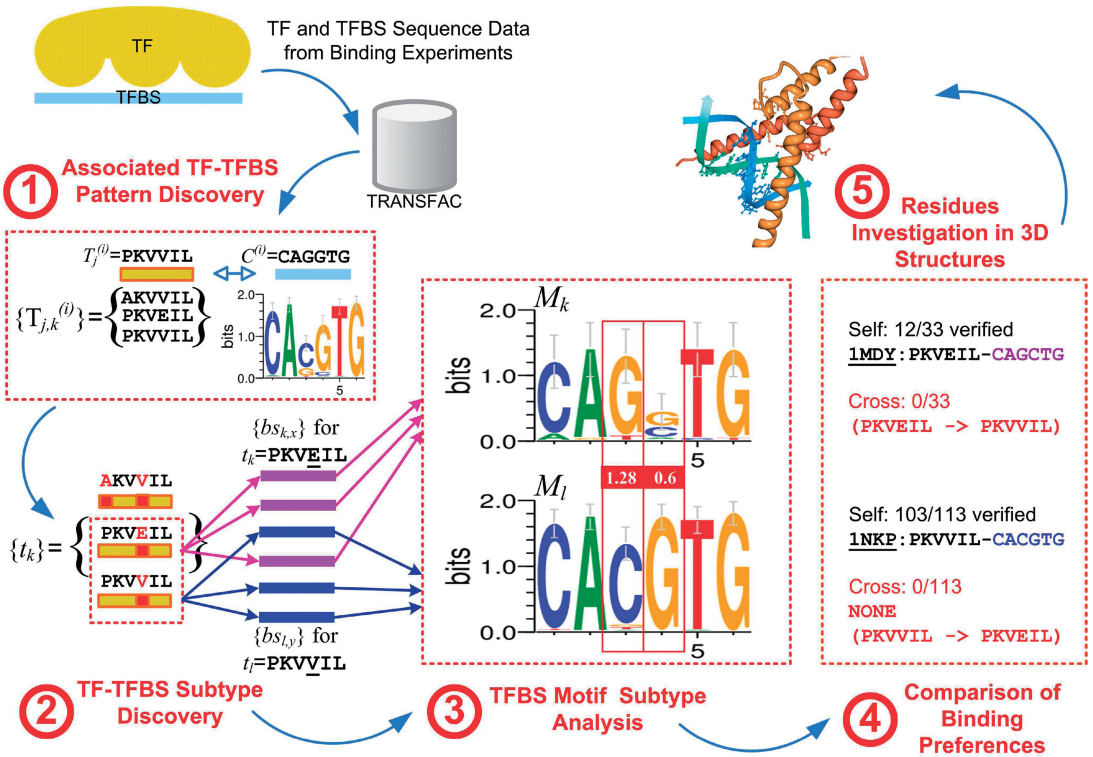

Figure 1. TF-TFBS subtype discovery flowchart. 1. Associated TF-TFBS pattern discovery (3). 2. TF-TFBS subtype discovery: t_k is the simplified notation for TF instance $T_{j,k}^{(i)}$; TFBSs are rearranged according to different t_k, e.g. PKVEIL and PKVVIL. 3. TFBS motif subtype analysis: the red bars in the middle indicate the statistically significant (P value ≤ 0.005) residue variations and the distances are shown in white. 4. Comparison of binding preferences: self: PDB verification ratio of the TFBS instances with the corresponding TF subtype; cross: verification ratio if the TFBS instances are associated with the other (opposite) TF subtype. 5. Residues investigation in 3D structures.

set at TY = 0.3 to encourage more diverse groups (3), the TF entries are grouped into 506 datasets each represented by the consensus identifier C_i. The approximate associated TF-TFBS pattern discovery is applied on all the data sets. The major parameters in the example are set as TF/TFBS width $W = 6,8$ and the maximal substitution error $E = 1,2$. The resultant patterns are in the form of $T_j^{(i)} - C^{(i)}$, where T and C represent the jth TF and the ith TFBS core patterns, i.e. PKVVIL and CAGGTG, respectively, in Figure 1, Part 1. In that running example, three different approximate instances belong to $T_j^{(i)}$ with mismatches/errors $\leq E$: **A**KVVIL, PKV**E**IL and PKVVI**L**, denoted as t_1, t_2 and t_3, respectively, when there is no ambiguity about i and j. They collectively form a set $\{t_k\}$.

The discovered associated patterns can be evaluated and verified using TF-TFBS 3D structures [1948 entries as in (3)] from PDB (9). We focus on the verification ratios, R_{TF} on the TF side and $R_{\text{TF-TFBS}}$ on both the TF-TFBS sides. For example, in Figure 1, Part 1, we have a pattern PKVVIL-CAGGTG. If under the pattern we have one AKVVIL, four PKVEIL and five PKVVIL TF instances, and both PKVEIL and PKVVIL can be found in core binding pairs in PDB, then $R_{\text{TF}} = (4+5)/(1+4+5) = 0.9$ for this associated pattern. If only PKVVIL-CACGTG (five instances) match the binding pairs in PDB on both sides while PKVEIL-CACGTG (four instances) does not, $R_{\text{TF-TFBS}} = 5/10 = 0.5$. Summarizing all associated patterns discovered, we can have the averaged AVG $R_{\text{TF-TFBS}}$ to measure the overall performance.

More technical details and the corresponding results can be found in the Supplementary Data.

TF-TFBS subtype discovery

On the basis of the successful discovery of approximate protein–DNA (TF-TFBS) associated patterns, we can further study the detailed binding variations contributed by residues which otherwise cannot be accommodated using the current limited amount of 3D structure data. By investigating the approximate (i.e. varying) TF instances in the associated patterns and extending the consensuses to the corresponding TFBS instances, we extract TF-TFBS residue variations (column-specific subtypes) and calculate the distances within a subtype pair. The details are elaborated as follows.

In the TF (core) motif discovery, data are subject to redundancy removal to avoid spurious motifs due to over-sampling (3). In residue variation subtype analysis, however, it is desirable to retrieve as many instances as possible, to ensure that the subtype discrimination does not happen due to insufficient data samples. Furthermore, examining samples beyond where the TF motifs are discovered can extensively verify the pattern generality. Therefore, distinct TF instances under a TF motif are retrieved from the whole TRANSFAC together with the corresponding TFBSs. As illustrated in Figure 1, Part 2, for each distinct instance t_k belonging to a TF motif, we search the whole TRANSFAC for the TF sequences containing it and retrieve all the corresponding bound TFBS instances, denoted by a set $\{bs_{k,x}\}$ where x is the index for a particular TFBS. Therefore, for two pairs $t_k - \{bs_{k,x}\}$ (PKVEIL-purple TFBSs) and $t_l - \{bs_{l,y}\}$ (PKVVIL-blue TFBSs), $t_k \neq t_l$, under the same associated pattern (PKVVIL-CAGGTG), we can investigate how the TFBSs $\{bs_{k,x}\}$ and $\{bs_{l,y}\}$ are different (i.e. TFBS subtypes) with respect to the TFBS consensus $C^{(i)}$.

As raw TFBSs have different widths, they are aligned to the consensus and truncated at the same width W. In particular, the W-conserved core of $C^{(i)}$ is chosen, i.e. the W subsequence of $C^{(i)}$ with the most conserved nucleotides. To measure nucleotide conservation in the consensus, we assign conservation score 1 for each conserved nucleotide A, C, G or T, 0.5 for mixed di-nucleotide R, Y, M, K, W or S, 0.33 for mixed tri-nucleotide B, D, H or V and 0.25 for the degenerated N. Then each TFBS (as well as the reverse complement) is aligned to the W-conserved core and the best aligned subsequence (substitution only) is extracted, resulting in the aligned core TFBS sets $\{bs_{k,x}\}$ and $\{bs_{l,y}\}$ with the same width W for two different TF instances t_k and t_l. Note that TRANSFAC contains noises and we discard TFBSs that can only poorly aligned with the consensus (mismatches $\geq 0.5 * W$). PWMs M_k for $\{bs_{k,x}\}$ (purple) and M_l for $\{bs_{l,y}\}$ (blue) are then generated accordingly as shown in Figure 1, Part 3.

TFBS subtype comparisons

To characterize and shortlist meaningful subtypes, comparisons are to be made between PWMs M_k (purple) and M_l (blue) corresponding to the two distinct TF motif instances t_k and t_l, respectively. In particular, column-wise Euclidean distance and the P values for Pearson chi-square test are employed. Although there are various comparison methods, Euclidean distance frequently shows best performance (17,19), and Pearson chi-square test is also among the top ones when sample size is $\geq$20 (19), which is exactly our case. Another reason to employ chi-square test is to easily obtain comparable statistics and intuitive P value thresholds to shortlist subtypes accommodating different sample sizes in different data sets. Other comparison methods such as Pearson Correlation Coefficient (PCC) and average log-likelihood ratio (ALLR) (17,19) produce consistent subtypes, mostly covered by our results with appropriate thresholds (see Supplementary Data for the comparisons). As the focused results are consistent, we employ P values and Euclidean distance in our analysis. Other advanced methods such as mutual information (MI) (21,22) could be employed in the future study.

The column-wise distance d_{col} can be calculated as follows:

$$d_{\text{col}}(M_k, M_l, q) = \left(\sum_b |M_k(b,q) - M_l(b,q)|^2\right)^{1/2} \quad (1)$$

where $b \in \{A,C,G,T\}$ is the nucleotide, q is the qth column in the PWM and L-2 norm (Euclidean distance) is employed in our experiments.

To investigate into residue varying subtypes, the column-wise distance d_{col} is focused on because the subtypes are largely similar while only small portions diverge. The positional discrimination is critical to

explain the binding differences for the highly conserved TF motif instances with only few mismatches.

Comparing the Euclidean distance (d_{col}) only may not be precise enough to evaluate subtypes, as there are sample size variations and biases in TRANSFAC leading to distorted distances. Therefore, statistical significance is employed to first compare and shortlist the TFBS PWM columns for two different TF instances t_k and t_l. Because we focus on residue varying subtypes, the statistical analysis is performed on the nucleotide level (PWM columns) rather than the motif level.

In particular, Pearson chi-square test of independence for two samples/outcomes (k and l) is employed. The null hypothesis is that the nucleotide frequency in the PWM column p (the occurrence of the outcomes) is statistically independent. The expected frequency for $r \in \{k,l\}$ and $b \in \{A,C,G,T\}$ is

$$E_{r,b} = \frac{(f_{k,b} + f_{l,b})(\sum_{c\in\{A,C,G,T\}} f_{r,c})}{N}, \quad (2)$$

where in general $f_{*,c}$ denotes the frequency count of nucleotide c in PWM column $M_*(c,q)$ for $* \in \{k,l\}$, and N is the total count of nucleotide frequencies of the two samples. The χ^2 statistic is

$$\chi^2(M_k, M_l, q) = \sum_{r\in\{k,l\}} \sum_{b\in\{A,C,G,T\}} \frac{(f_{r,b} - E_{r,b})^2}{E_{r,b}} \quad (3)$$

and the degree of freedom is (2–1)(4–1) = 3. The P value can be calculated at the resolutions of 0.001, 0.002, 0.005, 0.01, etc. The statistically significant subtypes are selected for further investigation based on the threshold P value ≤ 0.005. In other words, we only focus on those statistically significant subtype pairs in the text below. Corresponding results are in the later 'TF-TFBS Subtype Results' section.

Subtype investigation and analysis

There are various potential biological meanings and implications for the shortlisted TF-TFBS subtypes. They may reflect binding properties of the TF residues, exhibit changed binding preferences, differ in conservation due to contacting chemical bonds or carry other specific functions in regulatory mechanisms. The possibilities are investigated and verified with the following procedures from several aspects. The whole subtype discovery procedure is illustrated in Figure 1.

Overview

The subtypes imply different kinds of mechanisms and information with biological meanings:

- The subtypes may be directly involved in core TF-TFBS bindings. The varying residues may show flexibilities which are tolerated in the bindings on one hand and also exhibit changed binding preferences on the other. They are analyzed in sub-section 'Subtype Comparison of Binding Preferences' using 3D structures.
- The invariant (conserved) TF residues and/or TFBS nucleotides are likely to be critical contacting residues whose changes may significantly affect the chemical bonds. The information to discriminate conserved and flexible residues is useful for better modeling the error distributions of TF and/or TFBS motifs and improving existing motif representations, e.g. PWMs. They are elaborated through sub-section: 'Subtype Residues Investigation in 3D structures'.
- The subtypes may not be directly involved in protein–DNA contacting chemical bonds. However, the statistically significant variations are not likely to happen by chance. They may imply regulatory mechanisms and/or partners beyond direct protein–DNA interactions, for example, co-activator binding and dimerization. An interesting case that may lead to novel discoveries will be addressed in sub-section: 'In-depth Binding Analysis'.

Subtype comparison of binding preferences

To compare the binding preferences within the statistically significant subtype pairs, we make use of the procedure used in verifying approximate associated TF-TFBS patterns (3). The same binding protein–DNA (P-D) pairs from PDB are employed. To evaluate a TF subtype t_k (a distinct TF motif instance in width W) associated with a column from the corresponding set of aligned TFBSs ($\{bs_{k,q}\}$ in width W), we should evaluate t_k with each instance $bs_{k,q}$ through enumerating all $t_k - bs_{k,q}$ pairs. Each $t_k - bs_{k,q}$ is verified if both the TF and TFBS subsequences are contained in certain PDB P-D pairs. We record the verified count for the $t_k - bs_{k,q}$ pairs. Note that the verification is more stringent than our previous study (3), which only requires a TFBS consensus to be approximately matched for verification.

The next step is to investigate whether the residue-varying subtypes exhibit specific binding preferences by comparing them with artificial controls (TFBS subtypes associated with wrong TF instances). To check the binding preference of a pair of two TF subtypes t_k and t_l and their corresponding distinct TFBS patterns $bs_{k,q}$ and $bs_{k,r}$, we introduce cross-verification comparisons. Besides the previous PDB verification count on t_k with its actual corresponding TFBSs ($bs_{k,q}$), denoted as 'Self' verification, we perform the same PDB verification on the artificial control by associating the wrong t_l with $bs_{k,q}$, which is denoted as 'Cross' verification (the control). Note that the comparison is only performed when both TF instances t_k, t_l are verified on PDB, to avoid bias on TF instances with PDB evidence over those without. For the subtypes left out, we may resort to literature search and detailed 3D structure analysis.

Intuitively, 'Self' > 'Cross' indicates there are more PDB structure evidence verifying the correctly associated TF-TFBS subtype than the artificial control and thus supports existence of the binding preferences. All the 'Self' > 'Cross' percentages of the TF-TFBS subtypes are then reported. Besides 'Self' > 'Cross' verification performed on each 'individual' TF subtype with the binding preference ratio denoted as 'Individual Subtype Ratios',

binding preference ratios for subtype pairs are also introduced. If both TF subtypes in pair satisfy the 'individual' verification tests, they are said to satisfy the 'pair' verification test. The corresponding ratio is denoted as 'Pair Subtype Ratios'. One illustrative example is in Figure 1. Because both individual 'Self' > 'Cross' verification tests are satisfied for PKVEIL-CAGCTG (12/33 versus 0/33) and PKVVIL-CACGTG (103/113 versus 0/113), 'PKVEIL-CAGCTG versus PKVVIL-CACGTG' satisfies the 'pair' verification test. Corresponding results are in the 'Comparison Results of Binding Preferences on PDB' section.

Subtype residues investigation in 3D structures

As the previous comparisons are indirect and have the risk of sample bias in PDB, we further investigate into the detailed binding (interaction) properties of the varying and conserved residues on the PDB 3D structures. To generate a concise map from the many (redundant) subtypes, the statistically significant subtypes can be clustered according to the associated TF-TFBS patterns on both sides according to the maximal error E. As TFBSs are more flexible and less conserved, we focus on the residues (amino acids) on the TF side. They are categorized into two types in each cluster, the varying and the invariant (conserved) ones. It is then interesting to analyze the biochemical mechanisms of the strongly conserved residues (e.g. forming hydrogen bonds with the TFBS nucleotides) and the varying residues (e.g. showing specificities to different target TFBSs) by investigating the 3D structures one by one. Corresponding results are in the 'Results of Residues Investigation in 3D Structures' section.

In-depth binding analysis

Finally, for statistically significant subtypes without obvious evidence of direct interactions, we select interesting examples to perform literature search for reasonable interpretation. These cases are potential novel discoveries leading to co-factor bindings and revealing more intriguing regulatory mechanisms. Corresponding results are in the 'In-depth Analysis for Potential Co-factors' section.

RESULTS AND ANALYSIS

In this section, the detailed TF-TFBS subtype results and statistics are reported, followed by detailed variation analysis and comprehensive verification.

TF-TFBS subtype results

Based on the approximate associated TF-TFBS patterns discovered, subtype discovery was performed with width $W = 6,8$ and maximal error $E = 1,2$. All pairs of TF subtypes (i.e. different TF instances t_k and t_l within an associated TF-TFBS pattern) and their corresponding aligned TFBSs (PWMs M_k, M_l displayed as sequence logos in Figure 1) were analyzed. If there were any nucleotide differences leading to P value $P \leq 0.005$, the PWM column q associated with the TF subtypes are recorded and denoted as a subtype 'pair'. Meanwhile subtype pairs can be shortlisted by a column-wise distance threshold $C\text{dis} \geq d_{\text{col}}$. d_{col} is in the range of $[0, \sqrt{2}]$. As a proof of concept, we focus on the associated patterns with instance-level $R_{\text{TF-TFBS}} \geq 0.8$, and the same analysis can be applied on other patterns which are verifiable with more PDB records in the future (3). The statistics of the TF-TFBS subtypes are presented in Tables 1 and 2.

Subtype reflection of biological properties

We investigate into the TF subtype charge properties, which are summarized in Table 3 (details in the Supplementary Data). Residue variations are mainly within the same positive (Pos) or neutral (Neu) charge property groups, whereas variations with charge properties changed and variations within the negative (Neg) are rare. The observation is consistent with the chemical properties of protein–DNA interactions. As the backbone of DNA (TFBS) is negatively charged, positive and neutral residues being hydrophilic are more likely to be exposed on TF surfaces than negative ones. The DNA prefers positive/neutral amino acids to negative ones. Variations among different charge groups may significantly affect the bindings and thus are not preferred.

Table 1. The statistics of the TF-TFBS subtype pairs

Setting	Pair no.	Pattern no.	$C^{(i)}$ no.	d_{col}
W6E1	5108	643	145	0.22
W6E2	6343	463	158	0.26
W8E1	1250	182	69	0.28
W8E2	2262	175	67	0.29

Pair no. indicates the number of significant TF-TFBS subtype pairs with $P \leq 0.005$. Pattern no. indicates how many associated TF-TFBS patterns with $R_{\text{TF-TFBS}} \geq 0.8$ are included for the pairs. $C^{(i)}$ no. indicates the corresponding number of consensus groups (labeled by TRANSFAC PWM IDs). d_{col} is the average column-wise Euclidean distance.

Table 2. The accumulated TF-TFBS subtype counts with different Cdis values

$C\text{dis} \geq$	0.8	0.7	0.6	0.3	0.1
W6E1	5 (3)	39 (9)	88 (18)	947 (121)	4785 (145)
W6E2	41 (15)	102 (22)	214 (43)	1759 (140)	6057 (158)
W8E1	10 (7)	23 (8)	72 (13)	420 (50)	1239 (69)
W8E2	98 (8)	29 (13)	98 (22)	822 (57)	2237 (67)

The numbers of TRANSFAC TFBS consensus groups $C^{(i)}$ involved are shown in brackets.

Table 3. Summary of TF residue variation charge properties

Variation charge	Percentage	Chemical
Pos-Pos/Neu-Neu	High	Hydrophilic, preferred
Changed/Neg-Neg	Low	Not preferred

To focus on representative subtypes, we analyze the results with $Cdis \geq 0.6$ among of all different $Cdis$ (Table 2). On the subsets of $Cdis \geq 0.8,0.7,0.6$, we generated the Sequence Logos (23) for the TFBS patterns (PWMs) corresponding to different TF subtype pairs and visualized them for comparisons on the Project Website, as shown by the examples in Figures 2–4. In Figure 2a and b for $W = 6, E = 1$ and $Cdis \geq 0.6$, the G/V differences in the TF pair CKGFFK-CKVFFK lead to the different preferences of the 3rd and 4th nucleotides of the TFBSs. The TFBSs CKGFFK binds are more conserved at the two positions, whereas CKVFFK shows more flexibilities. Similar results can be observed for $W = 6, E = 2, Cdis \geq 0.6$ in Figure 3, indicating the consistency of the subtype discovery results. Figure 4 shows another related and consistent example, where C[DG/ES]CKGFF affects [A/C]AAGGTCA (illustrated in reverse complement TGACCTT[T/G].

As the subtypes were discovered without using any domain or family information, it is meaningful to check whether the subtypes reflect biological properties according to existing annotations. Therefore, for all the distinct TF subtypes with $Cdis \geq 0.6$, we checked the family annotations of their TF records in TRANSFAC (7), i.e. the class CL attribute in the factor records. We focused on TF subtypes with substantially different family information. As one short TF subtype instance can have many TF records with slightly different class annotations, we only recorded TF subtypes with different majority class (>50%) annotations. Interestingly, for $W = 6$, substantial subtypes, 62% ($E = 1$) and 80% ($E = 2$), respectively, show different major class annotations (note that subtypes were discovered without annotations).

For example, TFs with AKVVIL/PKVVIL and ERQRRN are all from bHLH-ZIP family (class C0012: bHLH-ZIP), whereas TFs with PKVEIL and ERRRRN are all from bHLH family (C0010: bHLH). The difference

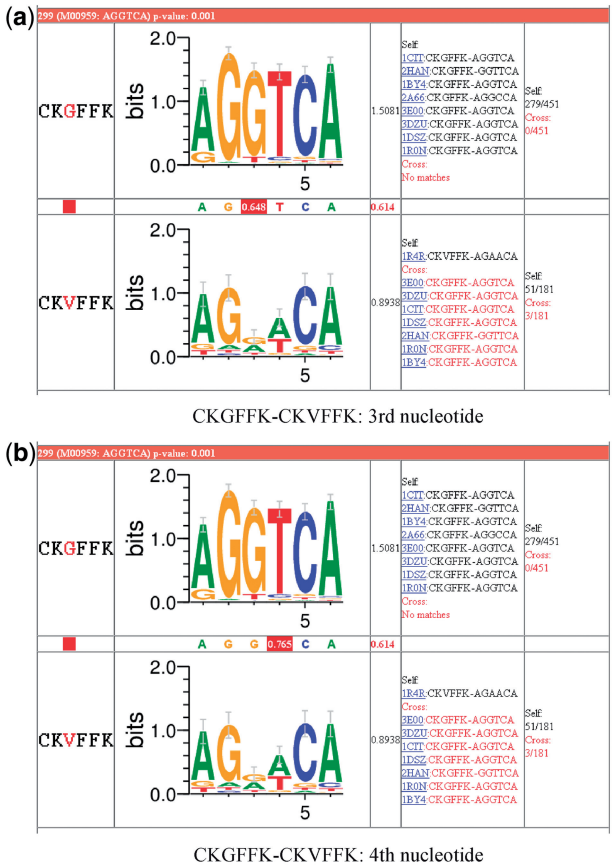

Figure 2. Subtype pair PDB verification and comparison for M00959: CKGFFK-CKVFFK, on the setting $W = 6$, $E = 1$. The variations on TF residues lead to the differences on 3rd ($Cdis = 0.648$) and 4th ($Cdis = 0.765$) nucleotides of the corresponding TFBS patterns (PWMs).

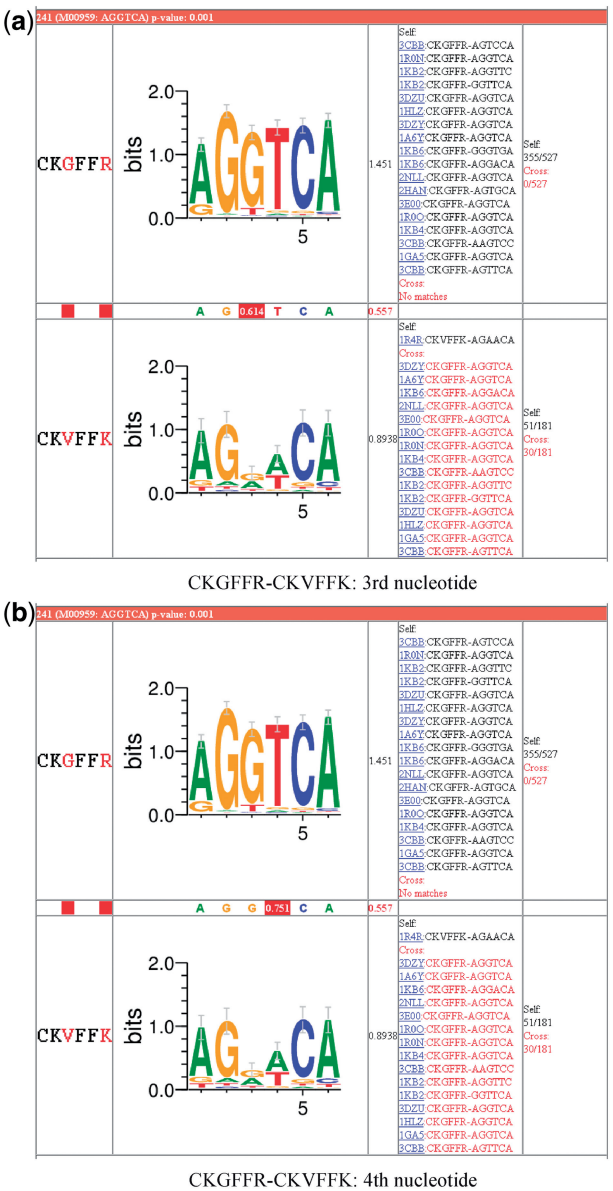

Figure 3. Subtype pair PDB verification and comparison for M00959: CKGFFR-CKVFFK, on the setting $W = 6$, $E = 2$. The variations on TF residues lead to the differences on 3rd ($Cdis = 0.614$) and 4th ($Cdis = 0.751$) nucleotides of the corresponding TFBS patterns (PWMs).

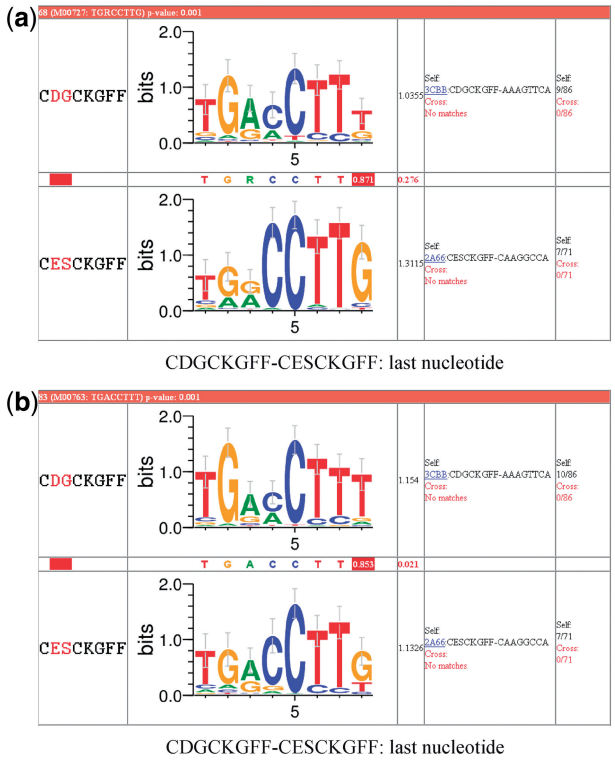

Figure 4. Subtype pair PDB verification and comparison for M00727: CDGCKGFF-CESCKGFF and M00763: CDGCKGFF-CESCKGFF, on the setting $W = 8$, $E = 1$. The variations on TF residues lead to the significant T/G changes in 8th position of the corresponding TFBS patterns (PWMs).

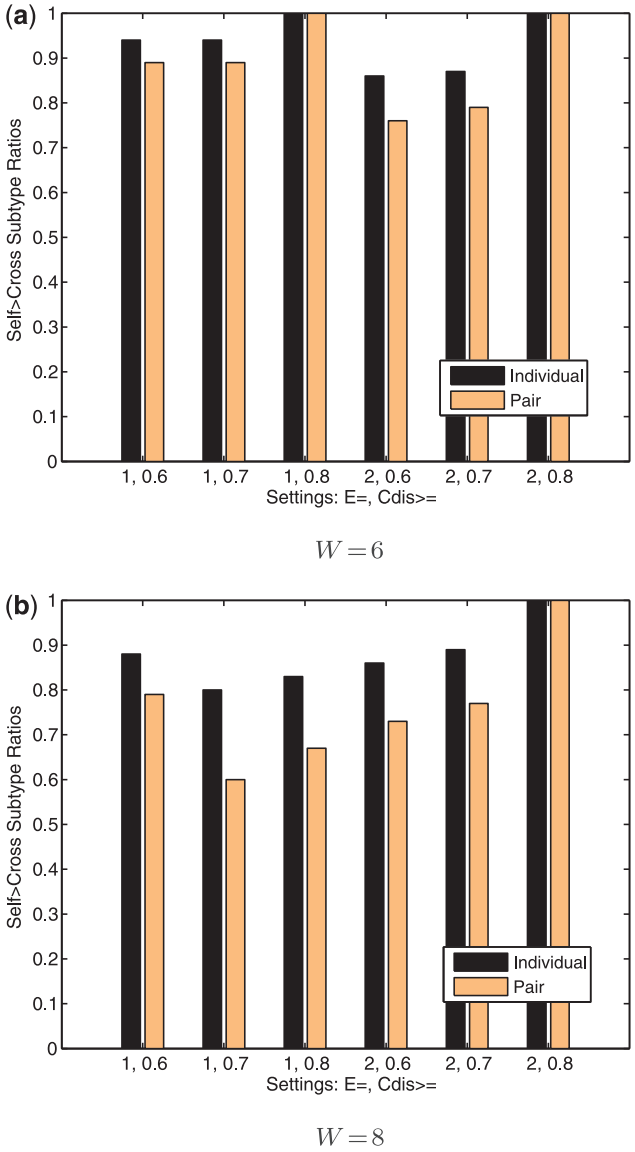

Figure 5. TF-TFBS 'Self' > 'Cross' verification subtype ratios on PDB records. Both individual and pair subtype ratios are shown. X axis indicates E and Cdis threshold, e.g. 1, 0.7 meaning $E = 1$, $C\text{dis} \geq 0.7$.

is whether there is a leucine zipper (ZIP) following the basic helix–loop–helix (HLH) domain. Another example is about subtypes of WFGNKR (C0006: homeo), WFQNAR (C0025: LIM-homeo) and WFQNHR (C0053: NK-2/Nkx). While the classes all belong to homeo box domains, the additional LIM motif in C0025 is a zinc-binding domain most likely involved in protein–protein contacts according to the class descriptions in TRANSFAC. NK-2/Nkx is the *Drosophila*-specific NK-homeobox to regulate gene transcription in a tissue- and developmental stage-specific manner. The annotation differentiation reflected by the subtypes not only shows family consistency with annotations but also reflects sub-familial functional specificities, including species, tissue and developmental specificities. Therefore, the subtypes are supported with biological meaningfulness.

For $W = 8$, the percentage drops to 23% ($E = 1$) and 24% ($E = 2$), respectively. The results imply that shorter subtypes ($W = 6$) tend to distinguish TF family classes, whereas longer subtypes ($W = 8$) tend to be variations within the same family or sub-family. This prompts for further interesting biological investigation into the subtypes in the future.

Comparison results of binding preferences on PDB

As mentioned previously, the 'Self' and 'Cross' verification counts on PDB were compared, where the latter represents a control scenario when the TFBSs belonging to a TF subtype are associated with the other ('counter-') TF subtype in the pair. The individual and pair subtype ratios on different Cdis threshold settings are shown in Figure 5 for both $W = 6$ and $W = 8$. The Self > Cross subtype ratios on $W = 6$ range from 0.86 to 1.00 for the individual ratios and 0.76 to 1.00 for the (more stringent) pair ratios. Similarly, the Self > Cross subtype ratios on $W = 8$ are in general ≥ 0.73 except for $E = 1$. The possible reason is that $W = 8$, $E = 1$ may be a too stringent approximation setting for approximate associated TF-TFBS subtypes, because the approximation level $E/W = 1/8$ is smaller than 1/6, 2/6 and 2/8. The results support that the TF subtypes do lead to specific binding preferences of TFBS pattern variations that are not likely interchangeable. These TF-TFBS subtypes provide more detailed information to recognize TFBS sub-patterns for these highly similar TF core subtypes and potentially better prediction

for TFBSs than mixing them together under the same pattern (PWM).

Support from ChIP-Seq motifs

Besides the PDB binding preference evaluations, high-quality motifs discovered from ChIP-Seq data can independently and indirectly verify our subtypes (discovered without ChIP-Seq). If one subtype (e.g. PKVVIL-CACGTG) from a subtype pair matches a ChIP-Seq motif well in terms of the binding TF type and the TFBS pattern (e.g. a similar motif logo), while the other subtype in the pair does not (e.g. KVEIL-CAG[G/C]TG), we consider the difference between the subtype individuals supported by ChIP-Seq evidence qualitatively. In other words, a subtype can better match the reliable ChIP-Seq motif evidence than an approximate pattern with all sub-patterns mixed.

We selected the two most conserved TFBS ChIP-Seq motifs from the recent stem cell study (8), namely the n/c-Myc and Esrrb (estrogen-related receptor beta) motifs discovered using both NMICA (24) and Weeder (25). We compared the two motifs with the most similar subtypes on the TFBS side for common properties. The examples, detailed in the Supplementary Data, show that our TFBS subtypes discovered are supported by the independent ChIP-Seq data with not only the TF family information but also the corresponding motifs.

Results of residues investigation in 3D structures

Besides the indirect binding preference comparisons, we investigated the varying and the invariant (highly conserved) TF residues in the clustered similar subtypes described in the methods section. We clustered the TF-TFBS subtypes with $W = 6$, $E = 2$ and $Cdis \geq 0.6$. In particular, associated TFs and TFBSs within $E = 2$ on both sides were grouped into one cluster. The columns with $Cdis \geq 0.6$ (when compared with any other instance) were marked by '*'. Note that different clusters may look like shifted versions with each other. Since merging shifted versions is non-trivial as the mismatches distributions will be changed and interpretations for merged clusters will be more involving, we will investigate it in future work. The full list is available on the Project Website. Here we focus on one of the clusters with at least four different TF-TFBS subtypes excluding the TFs without PDB binding-pair matches (labeled as N/A). One example cluster is shown below, in the form of $t_k - C^i$: PDB matches of t_k.

KAFFKR-AG**CA: 1HCQ; 1LAT; 1LO1;
KGFFKR-AG**CA: 1BY4; 1CIT; 1R0N; 1YNW;...
KGFFRR-AG**CA: 1A6Y; 1GA5; 1HLZ; 1KB2;...
KVFFKR-AG**CA: 1GLU; 1R4O; 1R4R; 2C7A;

In this example, if we consider the TF motif as KAFFKR, the TF instances are t_1 = KAFFKR, t_2 = KGFFKR, t_3 = KGFFRR and t_4 = KVFFKR (mismatches from the consensus are underlined). According to the residue categorization, K (1st), F (3rd), F (4th) and R (6th) are invariant residues, whereas the 2nd and 5th residues are the varying ones. The original TFBS consensus group $C^{(i)}$ is shown to be AGGTCA. The lists on the right show part of the PDB entries that verify t_k. We examined the corresponding PDB 3D structures (the first high-resolution one for each t_k) and checked how the invariant and varying residues participate in the protein–DNA bindings.

As illustrated in Figure 6, the interacting residues labeled are in general K's and R's. The 2nd varying residues (A/G/V) in the four different cases do not devastate the critical interactions but may contribute to a different binding TFBS subtype (V in 1R4R). Although the two invariant F's (phenylalanine's; labeled) do not directly interact with TFBS residues, they may be critical to maintain the suitable structure or architecture of the DNA binding domains in all the different TFs (1HCQ, 1BY4, 1A6Y and 1R4R) for the DNA to bind. If we look into the structure of the (helix turn helix) TFs, we always find two aromatic residues (phenylalanine's here) behind the DNA interacting residues (K's and R's here). The exceptions are the 5th residues (K/R), which are critical interacting and varying residues, but they both show similar properties such as being positively charged and highly hydrophilic.

The observation applies not only to this cluster but also to the other clusters (detailed analysis not shown). In the clusters, the residues interacting with the TFBSs are mostly positive charged, such as lysine (K) and arginine (R), and vary frequently. On the other hand, those

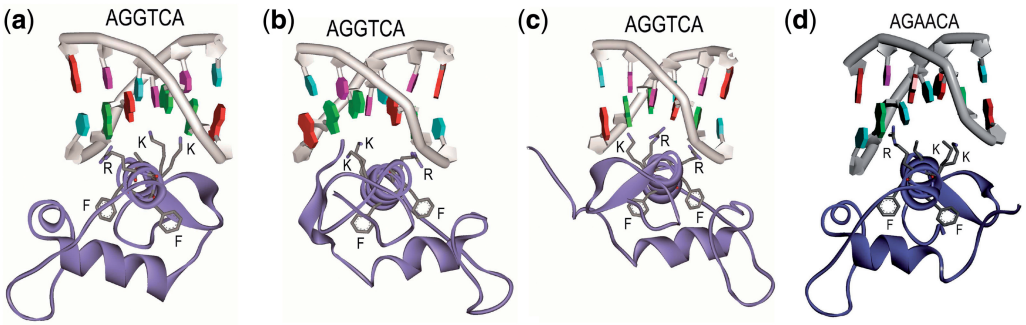

Figure 6. Three-dimensional investigation of a subtype cluster example on PDB. The important residues for the protein–DNA interactions are labeled.

residues in the middle (which are considered to maintain the structure of DNA binding domain) show some regular and more conserved patterns. If we also pay attention to the '*' ($Cdis \geq 0.6$) columns on the TFBS sides, we will find that in many cases the varying nucleotides are concentrated within certain columns. This is interesting and implies the subtypes on TF and TFBS sides may be from highly correlated and specific evolution (22).

In-depth analysis for potential co-factors

Apart from the previous residues investigation, here we analyze a typical subtype case without direct interaction evidence. The AKVVIL-CACGTG (No PDB)/ PKVVIL-CACGTG (verified in PDB 1NKP) (26) and PKVEIL-CAG[G/C]TG (verified in PDB 1MDY) (27) subtypes under 3D PDB investigation do not show clear residue-specific bindings (detailed in Supplementary Data). However, the corresponding TFBS patterns are highly different. Note that CACGTG and CAG[G/C]TG are not reverse complements of each other.

Nevertheless, we are interested in knowing the underlying mechanisms. The [A/P]KVVIL and PKVEIL subtypes are both found in the basic-helix-loop-helix (bHLH) domains, but the difference they exhibit is associated with different intriguing regulatory properties. By searching the 13 682 TF records of TRANSFAC, we find that the TF records with [A/P]KVVIL are all from the Myc (c-Myc) family (38 records), of 25 different species ranging from virus to human. Although AKVVIL mainly appears in c/N/S-Myc TFs (17 records), PKVVIL mainly appear in c/L/v-Myc TFs (21 records). On the other hand, TF records with PKVEIL are all from the myogenic regulatory factor (MRF) family (31 records), except one record named 'bHLH' lacking a specific TF name. In particular, PKVEIL appears in members of MRF including MRF4, Myf-5, Myf-6, MyoD, myogenin, Nau (nautilus) (28) and SUM-1 (sea urchin myogenic factor) (29), from 13 different species ranging from sea urchin to human. This is also supported by the PDB matches: the 1NKP (Myc-Max heterodimer) for PKVVIL-CACGTG and 1MDY (MyoD homodimer) for PKVEIL-CAGCTG.

The strong conservation and exclusiveness between [A/P]KVVIL and PKVEIL imply potential biological significance. Both the oncogenic Myc and the muscle-specific MyoD bind to a similar TFBS pattern called E-box (Enhancer Box) with consensus CANNTG, with a palindromic canonical sequence of CACGTG. However, previous research literature has discussed about the binding preferences of the binding patterns: CACGTG for Myc (30) and CAGCTG for MyoD (27), respectively.

We are interested in what roles the different V and E play in PKVVIL and PKVEIL, respectively. According to the crystal structures in 1NKP (26) with PKVVIL-CACG TG and literature search, Myc and Max form a heterodimer in DNA binding. Max lacks an activation segment and serves as an obligate physiological heterodimerization partner for Myc (26). Max has an R that contacts the N7 atom of the G (4th) of CACGTG (27). Myc-Max heterodimer further interacts with Miz-1 for repressing many genes (31,32). The binding capability can be lost with a few residue point mutations. In particular, if the V (4th) in PKVVIL of Myc is mutated to D, Myc fails to bind Miz-1 resulting in the binding deficient MycV394D (32). It is worth noting that both D and E exhibit similar properties: being polar, negative in charge and thus hydrophilic, whereas V (in [A/P]KVVIL) is non-polar, neutral and hydrophobic. Therefore, the PKVVIL and PKVEIL subtypes are not likely to be discovered just by chance. Moreover, the subtypes affect the co-factor binding and thus probably lead to different regulatory functions in the context of Myc Miz-1 bindings.

How the subtypes [A/P]KVVIL and PKVEIL in the bHLH family affect the regulatory mechanisms and whether they are related to the different TFBS subtypes of E-boxes are still open questions to us. Nevertheless, our study not only identifies TF-TFBS subtypes directly involved in binding preferences but also reveals subtype residues important for regulatory mechanisms through co-factor bindings. The V/E change is potentially important for categorizing the regulatory differences of the Myc oncoprotein and muscle-specific MRF family. More detailed regulatory mechanisms revealed for the Myc bindings will definitely help oncology study.

DISCUSSION AND CONCLUSION

In this study, we have for the first time introduced a large-scale subtype study based on the approximate associated protein–DNA (TF-TFBS) pattern discovery. Subtypes may lead to intriguing binding preferences and patterns, distinguish conserved (invariant) residues from flexible (varying) ones and reveal novel binding mechanisms. We have discovered subtypes of high statistical significance. Discovered without involving 3D structure experiments, the statistically significant TF-TFBS subtypes have exhibited intriguing binding preferences in verification comparison with PDB 3D structures and the examples have been supported by ChIP-Seq data. With more detailed 3D investigation on PDB structures, the example subtypes are shown highly indicative to distinguish the critical (invariant) residues and flexible (varying) residues on the TF side. The analysis also sheds light to potentially important residues for maintaining the structures/ architecture of the TF DNA binding domains. Further investigating a typical case without direct interaction evidence, we have found the TF subtypes are associated with regulatory mechanisms related to co-factor bindings. This study has identified more detailed TF-TFBS bindings associations, provided more solid and comprehensive evidence to support motif subtype discovery and complement the previous studies (21,22) on TF-TFBS binding co-evolution with limited data. Because of the limit of our manual analysis, there are still many more interesting subtypes to be analyzed in the next stage. They are publicly available. The study has shown potential to improve the understanding of gene regulation.

The subtype discovery is based on the approximate associated TF-TFBS pattern discovery, which makes use of existing TFBS consensus from TRANSFAC.

Further generalization on the modeling and search of associated TF-TFBS patterns would provide more informative and sound patterns to facilitate more powerful subtype discovery. One interesting direction is to set up a model that accommodates associated subtypes directly in TF-TFBS pattern discovery. The associated TF-TFBS subtype study with respect to various species is our next target. Further applications of the TF-TFBS subtypes include better prediction of TFBSs given the TF information based on a formal model and/or classifier and large-scale study on 3D structure data with the associated subtypes mined for more informative binding mechanisms.

The subtype discovery and analysis have broad potential biological applications. Subtypes can help to spot interesting variations that contribute to binding preferences, familial specificities and interacting mechanisms, as our comprehensive analysis has shown. Hierarchical annotation by introducing subtype besides site and domain is one possible extension (7,33). Follow-up phylogenetic studies on subtypes can further shed light on evolution and regulatory mechanisms. Subtypes can be utilized for more accurate motif discovery. Current motif matching suffers from false positives (34). As the associated TF-TFBS patterns and subtypes are further generalized, knowledge-driven motif matching can be developed as our next target to discriminate subtle variations to distinguish true TFBSs from false hits. With more and more data, subtypes can be further applied to analyze the mechanisms of alterations in the viral upstream regulatory region (URR) (35) to fight against diseases. The ultimate understanding TF-TFBS bindings will lead to artificial design of gene regulation circuits (36).

AVAILABILITY

Project Website: www.cse.cuhk.edu.hk/%7Etmchan/subtypes/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Text, Supplementary Tables 1–4 and Supplementary Figures 1–5.

ACKNOWLEDGEMENTS

The authors are grateful to the anonymous reviewers for their valuable comments.

FUNDING

Funding for open access change: The Focused Investment Scheme D on Hong Kong Bioinformatics Centre [Project No. 1904014]; Direct Grant [Project No. 2050500]; Chinese University of Hong Kong; GRF grant [Project No. 310111] from the Research Grants Council of Hong Kong SAR, China.

Conflict of interest statement. None declared.

REFERENCES

1. Luscombe,N.M. and Thornton,J.M. (2002) Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.*, **320**, 991–1009.
2. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, REVIEWS001.
3. Chan,T.-M., Wong,K.-C., Lee,K.-H., Wong,M.-H., Lau,C.-K., Tsui,S.K. and Leung,K.-S. (2011) Discovering approximate associated sequence patterns for protein-DNA interactions. *Bioinformatics*, **27**, 471–478.
4. Leung,K.-S., Wong,K.-C., Chan,T.-M., Wong,M.-H., Lee,K.-H., Lau,C.-K. and Tsui,S.K.W. (2010) Discovering protein-DNA binding sequence patterns using association rule mining. *Nucleic Acids Res.*, **38**, 6324–6337.
5. Smith,A.D., Sumazin,P., Das,D. and Zhang,M.Q. (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics*, **21(Suppl. 1)**, i403–i412.
6. MacIsaac,K.D. and Fraenkel,E. (2006) Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput. Biol.*, **2**, e36.
7. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, 108–110.
8. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
9. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
10. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (July, 2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.
11. 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.
12. Ofran,Y., Mysore,V. and Rost,B. (2007) Prediction of DNA-binding residues from sequence. *Bioinformatics*, **23**, i347–i353.
13. Bais,A.S.S., Kaminski,N. and Benos,P.V. (2011) Finding subtypes of transcription factor motif pairs with distinct regulatory roles. *Nucleic Acids Res.*, **39**, e76.
14. Li,M., Ma,B. and Wang,L. (2002) Finding similar regions in many sequences. *J. Comput. Syst. Sci.*, **65**, 73–96.
15. Chan,T.-M., Leung,K.-S. and Lee,K.-H. (2008) TFBS identification based on genetic algorithm with combined representations and adaptive post-processing. *Bioinformatics*, **24**, 341–349.
16. Chan,T.M., Li,G., Leung,K.S. and Lee,K.H. (2009) Discovering multiple realistic TFBS motifs based on a generalized model. *BMC Bioinformatics*, **10**, 22.
17. Mahony,S., Auron,P.E. and Benos,P.V. (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput. Biol.*, **17**, 578–591.
18. Habib,N., Kaplan,T., Margalit,H. and Friedman,N. (2008) A Novel Bayesian DNA motif comparison method for clustering and retrieval. *PLoS Comput. Biol.*, **4**, e1000010.
19. Gupta,S., Stamatoyannopoulos,J., Bailey,T. and Noble,W. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
20. Sarai,A. and Kono,H. (2005) Protein-DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.*, **34**, 379–398.
21. Mahony,S., Auron,P.E. and Benos,P.V. (2007) Inferring protein-DNA dependencies using motif alignments and mutual information. *Bioinformatics*, **23**, i297–i304.
22. Yang,S., Yalamanchili,H.K., Li,X., Yao,K.-M., Sham,P.C., Zhang,M.Q. and Wang,J. (2011) Correlated evolution of transcription factors and their binding sites. *Bioinformatics*, **27**, 2972–2978.

23. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
24. Doğruel,M., Down,T.A. and Hubbard,T.J.J. (2008) NestedMICA as an ab initio protein motif discovery tool. *BMC Bioinformatics*, **9**, 12.
25. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.
26. Nair,S.K. and Burley,S.K. (2003) X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors. *Cell*, **112**, 193–205.
27. Ma,P.C., Rould,M.A., Weintraub,H. and Pabo,C.O. (1994) Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. *Cell*, **77**, 451–459.
28. Michelson,A.M., Abmayr,S.M., Bate,M., Arias,A.M. and Maniatis,T. (1990) Expression of a MyoD family member prefigures muscle pattern in Drosophila embryos. *Genes Dev.*, **4**, 2086–2097.
29. Venuti,J.M., Goldberg,L., Chakraborty,T., Olson,E.N. and Klein,W.H. (1991) A myogenic factor from sea urchin embryos capable of programming muscle differentiation in mammalian cells. *Proc. Natl Acad. Sci.*, **88**, 6219–6223.
30. Blackwell,T.K., Kretzner,L., Blackwood,E.M., Eisenman,R.N. and Weintraub,H. (1990) Sequence-specific DNA binding by the c-Myc protein. *Science*, **250**, 1149–1151.
31. Wanzel,M., Herold,S. and Eilers,M. (2003) Transcriptional repression by Myc. *Trends Cell Biol.*, **13**, 146–150.
32. Si,J., Yu,X., Zhang,Y. and DeWille,J. (2010) Myc interacts with Max and Miz1 to repress C/EBPδ promoter activity and gene expression. *Mol. Cancer*, **9**, 1–5.
33. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., GrifRths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
34. Kel,A.E., Goessling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
35. Hubert,W.G. (2005) Variant upstream regulatory region sequences differentially regulate human papillomavirus type 16 DNA replication throughout the viral life cycle. *J. Virol.*, **79**, 5914–5922.
36. Zhan,S., Miller,J.F. and Tyrrell,A.M. (2009) An evolutionary system using development and artificial Genetic Regulatory Networks for electronic circuit design. *Biosystems*, **98**, 176–192.