BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 4 2011, pages 471–478
doi:10.1093/bioinformatics/btq682

Sequence analysis

Advance Access publication December 30, 2010

Discovering approximate-associated sequence patterns for protein–DNA interactions

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

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong, [2]Mathematical and Computer Sciences and Engineering Division, King Abdullah University of Science and Technology, Jeddah, KSA, [3]School of Biomedical Sciences, The Chinese University of Hong Kong and [4]Hong Kong Bioinformatics Centre, Shatin, N. T., Hong Kong

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The bindings between transcription factors (TFs) and transcription factor binding sites (TFBSs) are fundamental protein–DNA interactions in transcriptional regulation. Extensive efforts have been made to better understand the protein–DNA interactions. Recent mining on exact TF–TFBS-associated sequence patterns (rules) has shown great potentials and achieved very promising results. However, exact rules cannot handle variations in real data, resulting in limited informative rules. In this article, we generalize the exact rules to approximate ones for both TFs and TFBSs, which are essential for biological variations.

Results: A progressive approach is proposed to address the approximation to alleviate the computational requirements. Firstly, similar TFBSs are grouped from the available TF–TFBS data (TRANSFAC database). Secondly, approximate and highly conserved binding cores are discovered from TF sequences corresponding to each TFBS group. A customized algorithm is developed for the specific objective. We discover the approximate TF–TFBS rules by associating the grouped TFBS consensuses and TF cores. The rules discovered are evaluated by matching (verifying with) the actual protein–DNA binding pairs from Protein Data Bank (PDB) 3D structures. The approximate results exhibit many more verified rules and up to 300% better verification ratios than the exact ones. The customized algorithm achieves over 73% better verification ratios than traditional methods. Approximate rules (64–79%) are shown statistically significant. Detailed variation analysis and conservation verification on NCBI records demonstrate that the approximate rules reveal both the flexible and specific protein–DNA interactions accurately. The approximate TF–TFBS rules discovered show great generalized capability of exploring more informative binding rules.

Availability: Supplementary Data are available on *Bioinformatics* online and http://www.cse.cuhk.edu.hk/.

Contact: tmchan@cse.cuhk.edu.hk

Received on August 13, 2010; revised on October 31, 2010; accepted on December 7, 2010

*To whom correspondence should be addressed.

1 INTRODUCTION

Protein–DNA interactions in transcriptional regulation are first introduced, followed by the brief review on existing Bioinformatics methods to study protein–DNA interactions. The layout of the article is finally presented.

1.1 Protein–DNA interactions in transcriptional regulation

Protein–DNA interactions play a central role in genetic activities (Luscombe and Thornton, 2002; Luscombe *et al.*, 2000). The bindings of transcription factors (TFs) and transcription factor binding sites (TFBSs) are fundamental protein–DNA interactions in transcriptional regulation. Therefore, it is important to identify TF–TFBS binding rules to understand protein–DNA interactions and further decipher gene regulation. In particular, specific amino acids from the DNA binding domains of TFs can recognize and bind to similar short DNA binding sites (i.e. TFBSs, usually 5–20 bp) to regulate gene transcription. Based on functional similarities, the TF binding amino acids and TFBS have respective conserved patterns called motifs across different genes and/or species.

It is both expensive and time consuming to identify accurate TF–TFBS bindings experimentally either using the traditional DNA footprinting (Galas and Schmitz, 1987), gel electrophoresis (Garner and Revzin, 1981) or recent chromatin immunoprecipitation (ChIP) technology (MacIsaac and Fraenkel, 2006; Smith *et al.*, 2005). TRANSFAC (Matys *et al.*, 2006) is one of the largest and most representative databases for such regulatory elements including TFs, TFBSs, nucleotide distribution matrices of the TFBSs (TFBS motifs) and regulated genes. The data are annotated and curated from peer-reviewed and experimentally proved publications. Other annotation databases of TF families and binding domains are also available [e.g. PROSITE (Hulo *et al.*, 2008), Pfam (Bateman *et al.*, 2004)].

It is even more costly 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, which serve as valuable verification sources for putative binding discoveries. The Protein Data Bank (PDB) (Berman *et al.*, 2000) is the most representative repository with high resolution at atomic levels. However, the available 3D structures are far from complete. As a result, there is strong motivation to have automatic methods, particularly, computational approaches based

on other available data, to provide testable candidates of novel TF domains and/or TFBS motifs with high confidence to guide and accelerate the wet-lab experiments.

1.2 Existing bioinformatics methods

The first attempt of Bioinformatics methods to decipher TF–TFBS bindings was TF/TFBS motif discovery. Additionally, researchers have been trying hard for the protein–DNA one-to-one binding codes. Data mining methods have also been proposed, and recent work on mining exact TF–TFBS-associated sequence patterns shows promising results. They are briefly reviewed as follows:

Motif discovery: amino acids from TF domains and TFBSs sequences are conserved according to functional similarities. By exploiting conservation in the sequences, computational methods called motif discovery has achieved certain success in discovering TF or TFBS motifs. Motifs are usually represented as the consensus strings (Li *et al.*, 2002) or position weight matrices (PWMs) of the residue distributions (Stormo, 1988). *de novo* motif discovery (MacIsaac and Fraenkel, 2006) identifies the conserved patterns without knowing their motifs beforehand, based on certain motif models and scoring functions (Bailey and Elkan, 1994; Jensen *et al.*, 2004; Stormo, 1988) from a set of protein sequences/DNA promoter sequences with similar regulatory functions. A significant limitation of motif discovery is the lack of linkage between the binding counterparts and thus cannot directly reveal TF–TFBS relationships.

One-to-one binding codes: numerous studies have been carried out to analyze existing protein–DNA interaction structures comprehensively (Jones *et al.*, 1999; Krishna *et al.*, 2003; Luscombe and Thornton, 2002; Luscombe *et al.*, 2000, 2001). Various properties have been discovered concerning, e.g. bonding and force types, TF conservation and mutation (Luscombe and Thornton, 2002) and bending of the DNA (Jones *et al.*, 1999). Some are already applicable to predict binding amino acids on the TF side (Jones *et al.*, 2003). Alternatively, researchers have sought hard for general binding 'codes' between proteins and DNA, in particular the one-to-one mapping between the amino acids from TFs and the nucleotides from TFBSs. Despite many proposed one to-one binding propensity mappings (Luscombe and Thornton, 2002; Mandel-Gutfreund and Margalit, 1998; Mandel-Gutfreund *et al.*, 1995), it has come to a consensus that there are no simple binding 'codes' between single amino acids and nucleotides (Sarai and Kono, 2005).

Data mining: supervised learning approaches have also been proposed to mine protein–DNA interactions. Derived or transformed information is usually employed such as base compositions, structures, thermodynamic properties (Ahmad *et al.*, 2004, 2008) as well as expressions (Pham *et al.*, 2005). However, due to the stringent data requirement, many training based data mining methods concentrate only on specific families or particular datasets, and predicting only TF binding residues (Ahmad *et al.*, 2004, 2008), where the generality is limited. Furthermore, these methods usually produces predictors non-trivial to interpret (Ahmad *et al.*, 2004, 2008), and thus are less applicable for future general predictions.

On the other hand, sequences serve as the most handy and abundant primary data, and show promising results to reveal protein–DNA interaction relationships (Sarai and Kono, 2005). A recent association rule mining approach (Leung *et al.*, 2010) exploits the exact TF-TFBS-associated sequence patterns from TRANSFAC, and discovers informative rules verified on both literature and PDB structures. The study, however, is limited only on exact TF-TFBS-associated sequence patterns, while variations such as mutations and noises are common in real biological data. Moreover, the simple counts (supports), which can happen merely by chance, do not model overrepresentation biologically. As a result, the approach only generates a handful of exact rules, while there are still great potentials for many more flexible and verifiable rules to be discovered.

1.3 Paper layout

In this article, we generalize the exact TF–TFBS-associated sequence patterns to approximate ones on both sides. Many more informative rules are discovered for better understanding protein–DNA binding mechanisms. The article layout is as follows: the proposed methods are detailed Section 2; experimental results and verifications are reported in Section 3; and finally we have the Discussion and Conclusion in Section 4.

2 MATERIALS AND METHODS

In this section, we first present the overall framework for discovering approximate TF–TFBS rules, followed by the data preparation and detailed methodology.

2.1 Framework overview

To generalize exact TF–TFBS-associated sequence patterns (or rules for short) to approximate ones, direct modeling (scoring) TF–TFBS binding patterns as a whole is tempting but computationally challenging. To alleviate the difficulty, a progressive approach is proposed instead, as shown in Figure 1. Firstly, similar (approximate) TFBSs are grouped into a consensus C from the available TF–TFBS data (TRANSFAC database), and thus the binding TF sequences corresponding to group C form a TF dataset (with redundancy removed). Secondly, approximate and highly conserved motifs (cores) T are discovered from each TF dataset. The approximate TF–TFBS rules T-C are discovered by associating the TFBS consensuses C and the corresponding TF core motifs T. The detailed methodology is presented as follows.

2.2 Data preparation and TFBS grouping

To obtain the large-scale TF–TFBS binding sequence data, we employ the updated version of TRANSFAC Professional 2009.4 [an older public version (Matys *et al.*, 2006) is also available], which contains 13 682 TF entries (7664 with protein sequences) and 1225 matrices of the TFBS nucleotide distributions (TFBS motif matrices). Each TF is associated with the set of TFBSs it binds to, and matrices are the aligned and refined profiles of the similar TFBSs bound by the same TFs, with the motif consensus represented in IUPAC codes, which can be considered as the approximate TFBS motifs.

To simplify the TFBS grouping, we take advantage of the handy information of TFBS matrices (PWMs), in particular the TFBS motif consensuses, from TRANSFAC as part of the rules on the TFBS side. Note that the TFBS motif information is derived from TFBS sequence data using *de novo* motif discovery in TRANSFAC, so only TFBS sequence information is involved, and it is also possible to group raw TFBSs directly (more time consuming).

For each TFBS matrix, we use the IUPAC consensus as the TFBS motif, and cut all leading and ending 'N's (poorly conserved and non-informative). Similar motif consensuses are grouped with three different hamming distance ratio threshold TY's: 0.0, 0.1 and 0.3, reflecting different levels of approximation criteria. In particular, for each motif consensus C of the 1225 matrices from TRANSFAC, we align it (and its reverse complement)

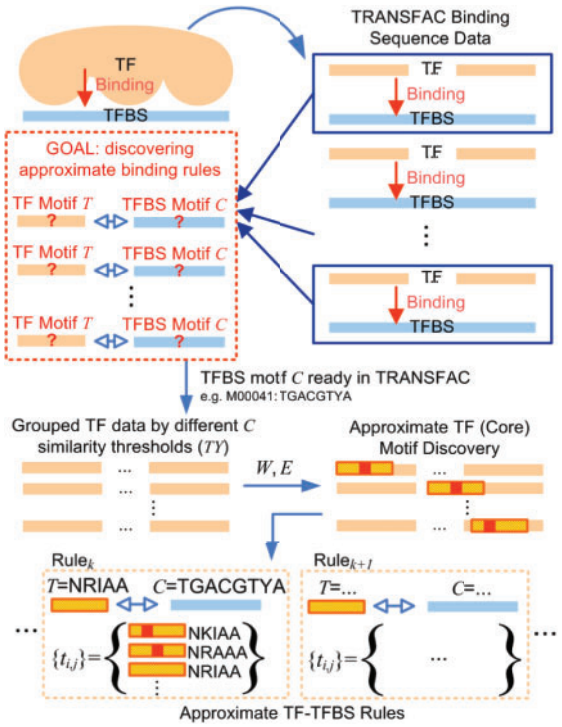

Fig. 1. The overall framework of approximate TF–TFBS rule discovery: C is the TFBS consensus group (IUPAC in TRANSFAC); T is the TF core with instance set $\{t_{i,j}\}$.

Table 1. The TF sequence datasets for different TFBS *TY*

TFBS *TY*	0.0	0.1	0.3
TF datasets	75 (475)	99 (490)	506 (815)

Numbers in parentheses: before redundancy removal.

with every other consensus C' for the best ungapped (substitution errors only) local pairwise alignment based on the hamming distance d. If d and the overlapping width w' between C and C' satisfy $d/w' \leq TY$, C' is grouped into C under threshold TY. Repeated consensuses are not processed again. For each TFBS consensus group, denoted by C, all the associated non-duplicate TF sequences are retrieved and then subject to CDHIT (with global sequence identify threshold -c 0.7) (Li and Godzik, 2006) to remove redundancy. Only non-redundant TF datasets with ≥ 5 sequences are kept. A summary of the TF datasets is shown in Table 1.

2.3 Approximate TF core motif discovery

Unlike the TFBS consensus groups, we have to find the potentially interacting motifs from the TF datasets (the middle part of Fig. 1). The core parts of TFs that closely interact (form bonds) with TFBSs are generally very short and do not vary largely due to their functional importance, so it is desirable to discover the short and conserved interacting amino acid subsequences from TFs. However, existing TF (protein) motif discovery methods [e.g. in MEME (Bailey and Elkan, 1994)] and annotation databases (Bateman *et al.*, 2004) mostly work on the domain level with low resolution, i.e. they aim at weakly conserved and long (≥ 30) motifs (Neduva and Russell, 2006). The few exceptions of short motif discovery methods, however, either eliminate domain-related subsequences (Neduva and Russell, 2006) that we need in our method or require non-trivial training only to discover underrepresented motifs (Doğruel *et al.*, 2008), which are not our targets. Furthermore, the general purpose methods neither explicit model conservation [e.g. TEIRESIAS (Rigoutsos and Floratos, 1998)] nor cater for binding favorable properties (Doğruel *et al.*, 2008; Neduva and Russell, 2006). Thus, we have to design a customized algorithm for the task, and useful features such as the hydrophilic properties favoring binding can also be incorporated.

The simple customized algorithm best fit our objective is described as follows. The inputs are the TF data with n sequences $S=\{S_i\}$, $i=1,...,n$ corresponding to a TFBS group C, the specified motif width W and the maximal error E. The outputs are the top K (=10 in our experiments) TF motifs T_k ($k=1,...,K$) and their corresponding matches $\{t_{i,j}\}_k$ maximizing certain motif scoring function f. i is the sequence index of S_i, and $j=0,1$ is the match index, indicating at most one match per sequence ($j=0$ means there is no match in S_i). Since the binding cores should be highly conserved, E is small in the expected target motifs. As a result, all W-substrings (W-mers) extracted by a sliding window on S are considered feasible to cover most of the probable motifs, without enumerating all 20^W possible W-mers. For each candidate motif T as a W-mer retrieved by the sliding window, all W-mers within hamming distance (substitution errors) E from T are retrieved as the candidate match set $\{tc_{i,j}\}$. i is the sequence index, and $j=1,...,q_i$, is the match index where q_i is the total number of matches in S_i. Exceptionally, $q_i=0$ means no candidate match for S_i. To favor the residues that are likely to be on the surface for binding, a candidate motif T should have at least one hydrophilic amino acid with a scale <0 (namely R, K, D, Q, N, E, H, S and T) from the normalized hydrophobic index (Eisenberg, 1984).

There can be several approximate matches to the same motif T from $\{tc_{i,j}\}$, but only the best match (one actual TF interacting core for one given TFBS core) should be chosen for each sequence. This is important but seldom considered by current pattern-based algorithms. Given the candidate set $\{tc_{i,j}\}$, we employ the Bayesian scoring function (Jensen and Liu, 2004) used for modeling conserved motifs to choose the most probable set of matches $\{t_{i,j}\}, j=0,1$ from $\{tc_{i,j}\}$. A customized iterative refinement approach is proposed. Firstly, all the first candidate matches, if any, are selected as the initial instance set $\{t'_{i,j}\} \leftarrow \{tc_{i,1}\}$ to build the initial PWM Θ of the amino acid distributions, where $\Theta_{a,b}$ represents the frequency of amino acid $b \in \Sigma$ at column $a \in [1,W]$. The background frequency of amino acid b, $\Theta_{0,b}$, can be calculated from input S. Then the Bayesian scoring function (Jensen and Liu, 2004) to be maximized is as follows:

$$f = |\{t'_{i,j}\}| \left(\sum_{a=1}^{w} \sum_{b \in \Sigma} \Theta_{a,b} \log \frac{\Theta_{a,b}}{\Theta_{0,b}} + \log \frac{p}{1-p} - 1 \right) \quad (1)$$

where $p = |\{t'_{i,j}\}|/|S|$ is the abundance ratio defined as the number of the matches, $|\{t'_{i,j}\}|$, over the dataset size $|S|$. The score reflects log posterior probability of having Θ and $\{t'_{i,j}\}$ with a non-informative prior. f can capture the overrepresentation and conservation concept of motifs with probability better than the simple supports (i.e. counts) (Leung *et al.*, 2010), which could be large by chance only.

The algorithm iteratively (maximal 20 iterations) tries the other candidates $tc_{i,j'}$ one by one at each S_i, and accepts the change if the new Θ improves f. If there is no change after trying all the matches from $\{tc_{i,j}\}$, the algorithm stops and outputs the top K best T associated with $\{t_{i,j}\}$. Utilizing the instance set with more stringent error constraints (E) has the advantage of being more concise for suppressing false positives. The instance set representation is also convenient for evaluation as shown later, because the ground-truth data are also arranged in an instance-based manner.

The algorithm converges very fast in experiments because there are only a few near-optimal matches to be chosen from each S_i with a small E set. To speed up, for each TF dataset, only the motifs with matches for $\geq n/2$ sequences are eligible to be processed to reduce computational time. Repeating motifs are not doubly processed. The approximate TF–TFBS rules T-C are finally formed by associating the TFBS consensus C and TF core motif T (instance set $\{t_{i,j}\}$).

3 RESULTS AND ANALYSIS

In this section, the discovered rules from experiments are reported, followed by detailed variation analysis and independent verification.

3.1 Experimental settings

With the 3 *TY* threshold settings of TFBS consensus grouping, different settings of $W=5,6$ and $E=0,1$ were used to run the TF motif discovery to generate different approximate TF–TFBS associated sequences patterns (referred simply as rules later on) from the extracted TRANSFAC data.

To evaluate the discovered rules based only on TF–TFBS sequences, the 3D protein–DNA complex structures from PDB were employed as the verification evidences. In particular, we downloaded 2457 PDB entries labeled with prot-nuc (protein–nucleotides) with redundancy removal at 90% sequence identity [same as the previous study (Leung *et al.*, 2010)]. We then removed entries without DNA chains (509 RNA entries), resulting in 1948 entries.

For each downloaded PDB entry, the distances between each amino acid on each protein chain and each nucleotide on each DNA chain were computed. If the respective residues (amino acid and nucleotide) have atoms that are close enough to be considered binding [$\leq$3.5 angstrom following (Ahmad *et al.*, 2004, 2008; Leung *et al.*, 2010)], the sequence pair *P-D* composed of the protein W'-mer *P* and DNA W'-mer *D* surrounding the particular close residues in the center was output, where W' is chosen as $2*W-1$. Thus, if a W'-mer contains a *W*-mer from the discovered rules, the *W*-mer is guaranteed to contain the close (binding) residue pair. Thus $W'=9,11$ for $W=5,6$ settings, respectively. These TF–TFBS W'-mer binding pairs (*P-D* pairs) were collected and compiled for the verifications (see Fig. 2). The summary is shown in Supplementary Table S1.

For each rule *T-C* specified by *W* (width only for TF, because *C* is retrieved from TRANSFAC) and error *E* with the TF instance set (optimal matches) $\{t_{i,j}\}$, there are two levels of PDB data verification: TF, verified on the TF side by protein (*P*) evidences, and TF–TFBS, verified on both sides by protein–DNA (*P-D*) evidences. To consistently compare with the previous study (Leung *et al.*, 2010), only rules with $\geq$7 instances are evaluated. The verification procedure is illustrated in Figure 2.

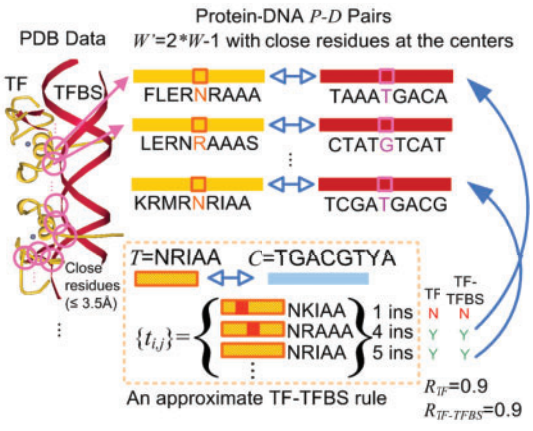

Fig. 2. An illustrative example of generating *P-D* pairs from PDB and verifying the approximate TF–TFBS rules for $W=5$, $E=1$ ($W'=9$). ins stands for TF instance(s).

TF side: since the instance set $\{t_{i,j}\}$ of the core motif *T* are obtained, one can directly check each TF instance $t_{i,j}$ for its presence in the protein substring *P* in PDB data. $t_{i,j}$ is verified on *P* if the *W*-mer $t_{i,j}$ is present in certain $W'=2*W-1$-mer(s) of *P* from the PDB *P-D* pairs, e.g. $t_{i,j}=NRAAA$ present in $P=FLERNRAAA$. The TF verification ratio R_{TF} for a rule is defined as the verified TF instance number over the total instance number $|\{t_{i,j}\}|$. Thus, if $E=0$, R_{TF} is either 0 or 1 because all instances are exact ones (the same). TF–TFBS sides: a TFBS motif consensus *C* from TRANSFAC is verified if there exist an *W*-mer in *C*, or its reverse complement, with at most *E* error from a present *W*-mer of *D* in the PDB *P-D* pairs. Note that since IUPAC code is employed in *C*, an ambiguity nucleotide can match any of its inclusive nucleotides (e.g. S matches C/G). For example with $W=5$ and $E=1$, $C=TGACGTYA$ is verified with $D=TCGATGACG$ because *TGACG* matches *D*'s last *W*-mer.

Thus, an approximate (*W*, *E*) TF–TFBS rule instance $t_{i,j}-C$ is verified if both the TF instance $t_{i,j}$ and the TFBS motif *C* can be verified on *P-D* PDB pairs. The TF–TFBS verification ratio $R_{TF-TFBS}$ for a rule *T-C* is defined as the verified $t_{i,j}-C$ number over the total rule instance number. Thus, $R_{TF-TFBS} \leq R_{TF}$. If $R_{TF}=0$ (not verified on TF side), $R_{TF-TFBS}=0$ (impossible to be further verified).

3.2 Approximate rule results

Table 2 shows the verification ratios, R_{TF} on the TF side and $R_{TF-TFBS}$ on both sides, on the corresponding PDB binding data, with respect to all TFBS consensus grouping *TY*, width *W* and error *E* settings. All detailed results of the rules are available in the Supplementary Material.

To compare with the previous study with exact TF–TFBS rules (Leung *et al.*, 2010), the results for $W=5,6$ (all rules with TF width *W* and TFBS width $\geq W$ are merged as one *W* setting for consistency) are collected and evaluated with the same verification procedures described above. The most exact setting from the approximate rules is $E=0$ for $TY=0.0$ (approximate information implicitly included in the IUPAC TFBS motifs).

The approximate rules have uniformly better average verified ratios (AVG R_*), e.g. better R_{TF} by 29% ($W=5$) and 300% ($W=6$), respectively, even when exact TF motifs are expected ($E=0$). Similar improvements on AVG $R_{TF-TFBS}$ are observed, with 46% ($W=5$) and 226% ($W=6$), respectively. It is expected because the exact rules are less favored with the limited TFBS widths discovered. The improved performance indicates the advantage of grouping approximate TFBS consensuses and discovering hydrophilic and probable TF motifs, over the exact counts (supports) (Leung *et al.*, 2010). Furthermore, with the approximate extensions, many more informative rules (rules with $R_*>0$) than exact ones are found, while maintaining competitive informative rule ratios ($R_*>0$ ratio). The previous exact rules (Leung *et al.*, 2010) become less appealing when *W* increases because there are fewer exact rules reaching the support threshold. Note that AVG R_* is equal to $R_*>0$ ratio when $E=0$ because all instances $t_{i,j}$ are the same and they are either 'all verified' ($R_*=1$) or 'none verified' ($R_*=0$) for a rule *T-C*.

The approximate rules also superset the exact ones in general. By summarizing all $E=0$ rules across different *TY* settings, the

Table 2. The verified rules on PDB binding data (*P-D* pairs) with different *TY*, *W* and *E* settings, compared with the corresponding $W=5,6$ exact rules in the previous study (Leung *et al.*, 2010)

	$W=5, E=0$		$W=5, E=0$						$W=5, E=1$					
TY	Exact rules (Leung *et al.*, 2010)		0.0		0.1		0.3		0.0		0.1		0.3	
R_*	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS
AVG R_*	0.57	0.44	0.74	0.64	0.78	0.70	0.82	0.73	0.57	0.56	0.63	0.62	0.69	0.68
$R_*>0$	99	76	127	110	165	147	636	567	235	231	291	287	2101	2072
Rule no.	173	173	172	172	211	211	774	774	346	346	396	396	2559	2559
$R_*>0$ Ratio	0.57	0.44	0.74	0.64	0.78	0.70	0.82	0.73	0.68	0.67	0.73	0.72	0.82	0.81
	$W=6, E=0$		$W=6, E=0$						$W=6, E=1$					
TY	Exact rules (Leung *et al.*, 2010)		0.0		0.1		0.3		0.0		0.1		0.3	
R_*	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS
AVG R_*	0.18	0.18	0.71	0.58	0.76	0.65	0.81	0.67	0.58	0.54	0.63	0.60	0.70	0.68
$R_*>0$	6	6	108	88	143	121	448	370	181	169	234	222	1665	1618
Rule no.	34	34	153	153	187	187	555	555	271	271	319	319	1920	1920
$R_*>0$ Ratio	0.18	0.18	0.71	0.58	0.76	0.65	0.81	0.67	0.67	0.62	0.73	0.70	0.87	0.84

R_* indicates R_{TF} or $R_{TF-TFBS}$.

approximate rules for $W=5, E=0$ cover 79% of the $W=5$ exact rules on TF sides and 79% on both sides. $W=5, E=1$ rules further cover 85% TF and 82% TF–TFBS exact rules. The small portions of the non-overlapping rules are probably due to the different data collection methods used [exact: TF oriented and all TFBSs used (Leung *et al.*, 2010); ours: TFBS consensus groups oriented and some original TFBSs ignored]. Approximate rules for $W=6, E=0$ also cover 88% TF and 85% TF–TFBS exact rules, respectively. Examples verified by the exact rules (Leung *et al.*, 2010) are also covered by the approximate rules. For example, the exact rule GGTCA-CEGCK, representing the P-box within Bp-nhr-2 binding domain (Moore and Devaney, 1999), is contained in 19 approximate rules (by matching the motifs) from all settings.

3.3 Comparisons with MEME on the TF motif discovery part

One may want to know whether traditional motif discovery methods can be incorporated in the TF motif discovery part of the whole TF–TFBS rule discovery. MEME, as a representative and widely used method, employs expectation maximization to discover motifs in the PWM representation, with minimal chance of having random motifs with better information content (IC) (Bailey and Elkan, 1994). Hence, MEME is likely to produce degenerate motifs (error *E* can be large). To check the suitability, we ran MEME on the same TF datasets and evaluated the final TF–TFBS rules in the same manner. MEME was set with fixed widths ($W=5,6$) and ZOOPS [zero or one (TF) instance per sequence] for consistency. AVG R_{TF}, AVG $R_{TF-TFBS}$ and $R_*>0$ Ratio were measured and compared with our approach. There is no error *E* parameter for MEME, so the same set of results for a specific *W* were measured twice with $E=0$ and $E=1$, of which the same R_{TF} results are expected because the TF performance measurement is instance oriented (matching $\{t_{i,j}\}$). On the other hand, $R_{TF-TFBS}$ will increase from $E=0$ to more relaxed $E=1$. The comparison results are shown in Table 3. Our approach is $73-262\%$ better in terms of AVG R_* than MEME for all different settings. MEME did find more rules in general because it tends to discover degenerate motifs. However, the verification ratios ($R_*>0$ Ratios) on all settings of our approach are $33-79\%$ better than MEME. Similar conclusions can be drawn from the comparisons with NMICA (Doğruel *et al.*, 2008), a recent method for short and degenerate protein motifs (see Supplementary Materials). The significantly better verification ratios of our method, designed specifically for highly conserved and short TF core motifs with hydrophilic constraints, indicate that it can better achieve the goal of discovering TF–TFBS rules than traditional motif discovery methods for general and degenerate motifs. Note that the experiments do not serve the purpose for selecting any better general-purpose motif discovery method, but to show our customized method is more suitable for this particular problem.

3.4 Statistical significance

To test the statistical significance ($W=5$ results for illustration) on R_{TF} and $R_{TF-TFBS}$, an empirical method is employed to simulate if the rules are randomly generated from the datasets. For each *TY* and *E* setting, each dataset corresponding to a TFBS consensus *C* is sampled equal times to output 10 TF motifs (denoted by T'), with *m* instances $t'_{i,j}$ generated with at most *E* from T', where *m* is randomly sampled to be valid for the above evaluation (i.e. ≥ 7 and $\geq n/2$, i.e. at least half of the sequence number). The sampling time for each *C* dataset is set such that there are $N \geq 10000$ datasets (e.g. $N = 134*75 = 10050$ for the 75 datasets with $TY=0.0$ and $E=0$) with totally $10*N$ rules generated. The empirical *P*-value of a rule is thus the proportion of random rules that has equal or better performance of R_* than it. The results for statistically significant rules (with $P<0.05$) for $W=5$ are summarized in Table 4. Note that for $E=0$, each random rule is either $R_*=0$ or $R_*=1$, and the best achievable *P*-values on TF side [i.e. $p(R_{TF}\geq 1)$] are 0.0625 (*TY*=0.0), 0.0668 (*TY*=0.1) and 0.0602 (*TY*=0.3). In such cases, the number of rules with the best achievable *P*-values are shown. It can be seen that the majority of the rules ($0.64-0.79$) are statistically significant for the TF–TFBS verification ratios $R_{TF-TFBS}$, indicating that the

Table 3. MEME results on different TY, W and E settings and the improved ratios of our approach over MEME (Ours better by referring to Table 2)

MEME results	$W=5, E=0$						$W=5, E=1$					
TY	0.0		0.1		0.3		0.0		0.1		0.3	
R_*	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS
AVG R_*	0.33	0.26	0.36	0.28	0.37	0.28	0.33	0.32	0.36	0.34	0.37	0.36
Ours better by (%)	124	144	120	146	120	160	73	74	76	79	85	91
$R_* > 0$	143	123	179	151	1306	1071	143	142	179	175	1306	1262
Rule no.	298	298	342	342	2118	2118	298	298	342	342	2118	2118
$R_* > 0$ Ratio	0.48	0.41	0.52	0.44	0.62	0.51	0.48	0.48	0.52	0.51	0.62	0.60
Ours better by (%)	54	55	49	58	33	45	42	40	40	42	33	36
MEME results	$W=6, E=0$						$W=6, E=1$					
TY	0.0		0.1		0.3		0.0		0.1		0.3	
R_*	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS
AVG R_*	0.29	0.22	0.31	0.23	0.29	0.18	0.29	0.27	0.31	0.29	0.29	0.26
Ours better by (%)	142	163	145	181	178	262	97	96	102	104	142	157
$R_* > 0$	127	96	163	121	1194	839	127	120	163	154	1194	1127
Rule no.	289	289	334	334	2170	2170	289	289	334	334	2170	2170
$R_* > 0$ Ratio	0.44	0.33	0.49	0.36	0.55	0.39	0.44	0.42	0.49	0.46	0.55	0.52
Ours better by (%)	61	73	57	79	47	72	52	50	50	51	58	62

R_* indicates R_{TF} or $R_{TF-TFBS}$.

Table 4. The statistically significant rules for $W=5$

	$W=5, E=0$						$W=5, E=1$					
TY	0.0		0.1		0.3		0.0		0.1		0.3	
R_*	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS	TF	TF–TFBS
$P < 0.05$	0 (127*)	110	0 (165*)	147	0 (636*)	567	223	226	278	272	1974	2023
Rule no.	172	172	211	211	774	774	346	346	396	396	2559	2559
Significant ratio	0 (0.74*)	0.64	0 (0.78*)	0.70	0 (0.82*)	0.73	0.64	0.65	0.70	0.69	0.77	0.79

*indicates the number of rules with the best achievable P-values when they are > 0.05 (all < 0.07).

competitive performances achieved by the approximate rules are not trivial.

3.5 Detailed analysis on variations

In this subsection, we investigate how the approximate rules generalize the exact ones using the verified PDB entries for illustration.

PDB examples and homology modeling: with the setting $W=5$ and $E=1$, we show how approximate rules generalize and retrieve informative verifiable evidences on both TF and TFBS sides. From the 231 verified ($R_{TF-TFBS} > 0$) TF–TFBS rules for $TY=0.0$, there are 133 verified rules with ≥ 5 PDB entries (maximum number of verified entries: 23). An illustrative rule with five verified PDB entries is chosen for illustration. The rule is M00041: NRIAA-TGACGTYA (ID 1160), with maximal $E=1$, the different TF instances (i.e. $\{t_{i,j}\}$) discovered by the customized algorithm are NKIAA, NRAAA, NREAA and NRIAA. Except NKIAA, other instances have been verified with PDB entries, namely 1DH3, 1FOS, 1JNM, 1T2K and 2H7H. The results are shown using ProteinWorkshop in Supplementary Figure S1. By allowing maximal 1 substitution error, we discover that the TF binding motif NR*AA summarized from our results is flexible with the middle amino acid, varying with E, A and I. Such discoveries supported by the approximate rules give us more clues into the TF–TFBS binding mechanisms.

In order to investigate the case of NKIAA, a model was built based on the structure of 1JNM using homology modeling. As shown in Figure 3, the change of arginine (R) to lysine (K) does not introduce the steric effect and the basic property of the amino acid is retained (both are positive charge). NKIAA is also shown to be within TF records of NCBI (Sayers *et al.*, 2010) in the next subsection. Thus, we believe that NKIAA should be a correct prediction.

We further analyze the rule picked up from setting $W=5$, $E=1$ and $TY=0.1$. The rule M00217: ERKRR-CACGTG has three different TF instances (i.e. $\{t_{i,j}\}$) ERKRR, ERQRR and ERRRR, and five verified PDB entries: 1AN2, 1AN4, 1HLO (not shown for space limit), 1NKP and 1NLW. The results are shown using ProteinWorkshop in Figure 4. This case further demonstrates the flexibility in specific positions for TF–TFBS binding. ER*RR has the

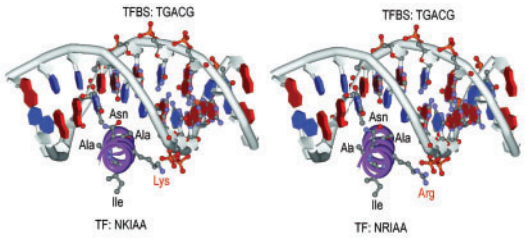

Fig. 3. Homology modeling of NKIAA-TGACG that does not have PDB records, based on the verified NRIAA-TGACG pair. The model (left) was built based on and compared with the structure of 1JNM (right). The proteins are shown in ribbon diagram with the highlighted TF amino acids in ball and stick format. The TFBS sequences in the DNA are also highlighted in ball and stick format. The figures are generated using Discovery Studio Visualizer, Accelrys.

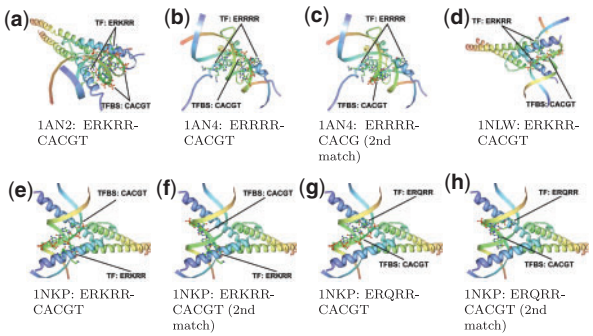

Fig. 4. PDB verifications for rule M00217: ERKRR(ERKRR; ERQRR; ERRRR)-CACGTG for $W=5$, $E=1$, $TY=0.1$ using ProteinWorkshop. The TF–TFBS pairs are shown in ribbon diagram and labeled. The interacting TF amino acids and TFBS nucleotides are highlighted in ball and stick format.

variations of K, R and Q for the middle amino acid, and these variants can appear in the same TF–TFBS binding, for example, 1NKP (ERKRR and ERQRR) in Figure 4. The discovery prompts further investigation into the flexibility and specificity of protein–DNA interactions.

In-depth variation analysis: detailed analysis was also performed on the properties of residue variations reflected by the approximate rules. As a proof of concept, the 134 rules with $R_{TF-TFBS} \geq 0.9$ from setting $W=5, E=1, TY=0.1$ were investigated. Excluding the 39 rules with only exact TF instances (i.e. no variations) according to the PDB evidences, we checked the varying amino acids of the remaining 95 approximate rules (i.e. $\{t_{i,j}\}$ with $E=1$). Interestingly, 50 rules (52.36%) contain varying amino acids that do not affect the binding (with distances to nucleotides >3.5 Å), and thus the approximate rules correctly reflect the flexibilities of such residues. For example, the varying residue (I or V) in the concatenated TF motif V[I/V]RVWFCN is flexible because it is not directly interacting with the nucleotides. On the other hand, 45 rules (47.37%) contain some TF instances with varying amino acids $\leq$ 3.5 Å to nucleotides, e.g. N[K/R]IAA and ER[K/R/Q]RR mentioned above, and F[Q/R][I/V]PW[K/M]H[A/F/G]-RAAANTGAAA. The last residue (A or F or G) of the latter example is varying but still interacting with the nucleotides ($\leq$3.5). Interestingly, the interacting variations in 32 of the 45 rules (71.11%) are consistent with respect to hydropathy [the varying residues are either all-hydrophilic or all-hydrophobic according to the normalized hydrophobic index (Eisenberg, 1984)], implying the biological clues of the variations. The remaining 13 rules contain variations between hydrophilic and hydrophobic amino acids, and the 11 ambiguous C/Q variation rules can be discarded because the C variants do not affect binding (>3.5 Å), while the Q variants do, in the PDB records. For the remaining two rules, the residue variations come with variations in the respective nucleotides of the TFBS they bind, for example, in the rule SG[F/K/Y]HY-TGACCTTTGNCCY (reverse complement RGGNCAAAGGTCA), when Y $\rightarrow$ K, the TFBS *CAAGGT $\rightarrow$ *AAAGGT; when F $\rightarrow$ K, *TAGGT $\rightarrow$ *AAGGT, demonstrating the coordination needed to cater for property changes in the binding amino acids. This preliminary discovery prompts us to deeply investigate the flexibility and specificity of protein–DNA interactions in the future work, with the help of approximate TF–TFBS rules.

3.6 Conservation verification on NCBI protein records

Besides the PDB entries, we further verified the approximate rules on NCBI (Sayers *et al.*, 2010) for conservation independently. The previous 134 rules with $R_{TF-TFBS} \geq 0.9$ ($W=5, E=1, TY=0.1$) were compiled (grouped) according to their 39 different TFBS consensus C groups, and the first 10 groups were analyzed for illustration (because of the time-consuming manual inspection). For each C, the TF names FA and organisms OS of the related TFs were retrieved, and TF instances ($\{t_{i,j}\}$) found in the approximate rules were recorded. We then queried proteins in NCBI with FA, and check whether any instance in $\{t_{i,j}\}$ occurs in protein records of organisms **NOT included in** OS.

All the 10 groups are conserved within protein records in NCBI from organisms not recorded in the TRANSFAC data (see Supplementary Material for details). All of the TF instances are within the conserved domains (especially binding domains), except one case where the domain information is missing in NCBI, and overlap with the annotated DNA binding sites. For example, NREAA, NRAAA in the 1st, 7th and 10th groups are conserved among proteins (TFs) CREB1, ATF-1 in various organisms such as *Danio rerio*, *Oncorhynchus mykiss* and *Saccharomyces cerevisiae*, which are beyond the TRANSFAC data containing mainly higher mammals. None of these organisms are included in the corresponding TRANSFAC data used to discover the rules. Furthermore, the conserved TF instances are all within consistent conserved domains and overlapping with binding sites according to the NCBI annotations. For example, the conserved ERQRR and ERRRR from the 6th group are all within helix–loop–helix (HLH) domains in NCBI although they appear in various proteins such as USF, N-Myc and arnt. The confirmation of conservation of the discovered TF instances in NCBI records strongly indicates that the approximate TF motifs are very likely to be real conserved binding cores across different organisms (especially when they are within consistent conserved domains and overlapping with DNA binding sites), thus demonstrating the accuracy and generality of the approximate rules for revealing real TF–TFBS interactions.

4 DISCUSSION AND CONCLUSION

Data mining on sequence patterns from large-scale databases shows great potentials for discovering TF–TFBS rules for further

understanding protein–DNA interactions. In this article, we have for the first time generalized the exact TF–TFBS rules (Leung *et al.*, 2010) to approximate ones to discover more informative and intricate rules. Reliable datasets are ready for use through grouping the non-redundant TF sequences corresponding to similar TFBS consensuses C, which has greatly accelerated the study. A simple customized algorithm has been developed to discover the short (width $W = 5, 6$) and highly conserved (error $E = 0, 1$) TF core motifs. The algorithm better suits our objective and significantly outperforms MEME by over 73%. Comprehensive measures, e.g. both TF and TF–TFBS verification ratios (R_*), verified rule ratios ($R_* > 0$ Ratios), as well as statistical significances have been used to evaluate the discovered approximate TF–TFBS rules.

The discovered approximate TF–TFBS rules have demonstrated competitive performance with respect to verifications ratios (R_*) on both the TF and the TF–TFBS sides. The approximate rules exhibit a strong edge over the previous exact ones on both average verification ratios and number of informative rules, where the majority are shown to be statistically significant. With detailed analysis, the approximate rules are confirmed by the PDB binding structures visually and interatomic distances computed, as well as homology modeling for the rule without PDB records. The examples have demonstrated the flexibility of specific positions with variations of TF–TFBS binding for both proteins and DNAs, reinforcing the need to extend exact rules to approximate ones to better discover TF–TFBS binding patterns. The approximate TF instances corresponding to the rules discovered are conserved in binding domains and even binding sites according to the independent verification on NCBI records from organisms not included in the TRANSFAC data used, and hence strongly support the biological significance of the discovered rules.

Compared with the previous study on exact rules, the proposed discovery of approximate TF–TFBS rules has demonstrated significantly better generalized capability of exploring more informative binding rules, and potential applications to predict protein–DNA interactions given either side for better decipher transcriptional regulation. As advanced computational techniques, facilities and databases grow rapidly, there will be numerous promising ways to further improve approximate TF–TFBS rule discovery greatly. One promising direction is expressive probabilistic representations, such as Hidden Markov Models (Bateman *et al.*, 2004), which are able to capture subtle information and dependencies for long TF–TFBS rules.

ACKNOWLEDGEMENTS

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

Funding: The research is supported by the grant CUHK414708 from the Research Grants Council of the Hong Kong SAR, China, and Focused Investment Scheme D on Hong Kong Bioinformatics Centre (Project Number: 1904014) from The Chinese University of Hong Kong.

Conflict of Interest: none declared.

REFERENCES

Ahmad,S. *et al.* (2004) Analysis and prediction of dna-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.

Ahmad,S. *et al.* (2008) Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Res.*, **36**, 5922–5932.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, AAI press, pp. 28–36.

Bateman,A. *et al.* (2004) The pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Doğruel,M. *et al.* (2008) Nestedmica as an ab initio protein motif discovery tool. *BMC Bioinformatics*, **9**, 19.

Eisenberg,D. (1984) Three-dimensional structure of membrane and surface proteins. *Annu. Rev. Biochem.*, **53**, 595–623.

Galas,D.J. and Schmitz,A. (1987) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.*, **5**, 3157–3170.

Garner,M.M. and Revzin,A. (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the escherichia coli lactose operon regulatory system. *Nucleic Acids Res.*, **9**, 3047–3060.

Hulo,N. *et al.* (2008) The 20 years of prosite. *Nucleic Acids Res.*, **36**(Suppl. 1), D245–D249.

Jensen,S.T. and Liu,J.S. (2004) BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics*, **20**, 1557–1564.

Jensen,S.T. *et al.* (2004) Computational discovery of gene regulatory binding motifs: a bayesian perspective. *Stat. Sci.*, **19**, 188–204.

Jones,S. *et al.* (1999) Protein-dna interactions: a structural analysis. *J. Mol. Biol.*, **287**, 877–896.

Jones,S. *et al.* (2003) Using electrostatic potentials to predict dna-binding sites on dna-binding proteins. *Nucleic Acids Res.*, **31**, 7189–7198.

Krishna,S.S. *et al.* (2003) Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.*, **31**, 532–550.

Leung,K.-S. *et al.* (2010) Discovering protein-DNA binding sequence patterns using association rule mining. *Nucleic Acids Res*, **38**, 6324–6337.

Li,M. *et al.* (2002) Finding similar regions in many sequences. *J. Comput. Syst. Sci.*, **65**, 73–96.

Li,W. and Godzik,A. (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

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.

Luscombe,N.M. *et al.* (2000) An overview of the structures of protein-dna complexes. *Genome Biol.*, **1**, REVIEWS001.

Luscombe,N.M. *et al.* (2001) Amino acid-base interactions: a three-dimensional analysis of protein-dna interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.

MacIsaac,K.D. and Fraenkel,E. (2006) Practical strategies for discovering regulatory dna sequence motifs. *PLoS Comput. Biol.*, **2**, e36.

Mandel-Gutfreund,Y. and Margalit,H. (1998) Quantitative parameters for amino acid-base interaction: implications for prediction of protein-dna binding sites. *Nucleic Acids Res.*, **26**, 2306–2312.

Mandel-Gutfreund,Y. *et al.* (1995) Comprehensive analysis of hydrogen bonds in regulatory protein dna-complexes: in search of common principles. *J. Mol. Biol.*, **253**, 370–382.

Matys,V. *et al.* (2006) Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, 108–110.

Moore,J. and Devaney,E. (1999) Cloning and characterization of two nuclear receptors from the filarial nematode Brugia pahangi. *Biochem. J.*, **344** (Pt 1), 245–252.

Neduva,V. and Russell,R.B. (2006) Dilimot: discovery of linear motifs in proteins. *Nucleic Acids Res.*, **34** (Suppl. 2), W350–W355.

Pham,T.H. *et al.* (2005) Computational discovery of transcriptional regulatory rules. *Bioinformatics*, **21**, 101–107.

Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the teiresias algorithm. *Bioinformatics*, **14**, 55–67.

Sarai,A. and Kono,H. (2005) Protein-dna recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.*, **34**, 379–398.

Sayers,E.W. *et al.* (2010) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **38**, D5–D16.

Smith,A.D. *et al.* (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics*, **21** (Suppl. 1), i403–i412.

Stormo,G.D. (1988) Computer methods for analyzing sequence recognition of nucleic acids. *Annu. Rev. BioChem.*, **17**, 241–263.