Modeling Associated Protein-DNA Pattern Discovery with Unified Scores

Tak-Ming Chan, Leung-Yau Lo, Ho-Yin Sze-To,
Kwong-Sak Leung, Xinshu Xiao, and Man-Hon Wong

Abstract—Understanding protein-DNA interactions, specifically transcription factor (TF) and transcription factor binding site (TFBS) bindings, is crucial in deciphering gene regulation. The recent associated TF-TFBS pattern discovery combines one-sided motif discovery on both the TF and the TFBS sides. Using sequences only, it identifies the short protein-DNA binding cores available only in high-resolution 3D structures. The discovered patterns lead to promising subtype and disease analysis applications. While the related studies use either association rule mining or existing TFBS annotations, none has proposed any formal unified (both-sided) model to prioritize the top verifiable associated patterns. We propose the unified scores and develop an effective pipeline for associated TF-TFBS pattern discovery. Our stringent instance-level evaluations show that the patterns with the top unified scores match with the binding cores in 3D structures considerably better than the previous works, where up to 90 percent of the top 20 scored patterns are verified. We also introduce extended verification from literature surveys, where the high unified scores correspond to even higher verification percentage. The top scored patterns are confirmed to match the known WRKY binding cores with no available 3D structures and agree well with the top binding affinities of in vivo experiments.

Index Terms—Bioinformatics, protein-DNA interactions, motif discovery, TF-TFBS associated pattern discovery, binding rules

1 Introduction

Protein-DNA interactions play a fundamental and essential role in various genetic activities [1], [2]. The proteins called transcription factors (TFs) recognize and bind to short DNA regions called transcription factor binding sites (TFBSs) in a sequence specific manner, and activate or suppress the target gene expression. As the primary protein-DNA interactions in gene regulation, TF-TFBS bindings will be our focus throughout the paper. There are numerous studies to decipher their patterns as a critical component in understanding life and disease mechanisms for bioengineering and therapeutic purposes [1], [3], [4], [5].

The protein-DNA interactions and their conserved patterns can be categorized in different sequence resolutions. There are substantial differences across full-length TF sequences with lengths of hundreds of residues (amino acids) and experimentally determined bound TFBS sequences with lengths within twenty residues (nucleotides or base pairs bp). Despite the remarkable global sequence differences, shorter consecutive regions of the TF sequences are conserved and form binding domains, resulting in a significantly smaller number of characteristic families [6], [7]. These binding domains, with lengths of tens to around one hundred residues, have similar 3D local structures and sequence specific preferences for binding to TFBS patterns of around several to 20 bp, which are called motifs. Even shorter TF and TFBS subsequences critical for bindings can be extracted from experimentally determined 3D protein-DNA complex structures. In particular, atomic distances can be measured between all pairs of TF and TFBS residues, and the pairs with atoms ≤ 3.5 Å [8], [9] are considered forming chemical (hydrogen) bonds of the bindings. We denote the TF-TFBS subsequences with lengths (widths) of several residues surrounding the bonding pairs as binding cores [10], which can be considered as the sequence representation of the interaction interfaces [11]. *The binding cores are the most critical parts of protein-DNA interactions and are much shorter than the sequence lengths of the whole 3D structures in the TF domain scale.*

Identifying binding cores on both the TF and TFBS sides is challenging experimentally. It requires resolving the protein-DNA complex 3D structures that are deposited and available in the protein data bank (PDB) [12]. Available protein-DNA 3D structures are limited and far from complete because of the experiment costs and difficulties. For example, there is an over 10-time sequence-structure gap regarding the number of known protein sequences and that of known protein structures in PDB, and moreover, many structures in PDB contain only single protein domains but not the protein-DNA interaction complexes, with no information of the binding cores (interfaces) [13].

On the other hand, experiments determining TF-TFBS bindings at the sequence level are inexpensive with abundant data. High-throughput experiments, such as in vivo Chromatin immunoprecipitation followed by sequencing (ChIP-seq) [14], [15], [16], and in vitro protein binding

- *T.-M. Chan is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong, and the Department of Integrative Biology and Physiology, University of California Los Angeles, 611 Charles E. Young Drive, Los Angeles, CA 90095-1570. E-mail: tmchan@cse.cuhk.edu.hk, cyruschan@ucla.edu.*
- *L.-Y. Lo, H.-Y. Sze-To, K.-S. Leung, and M.-H. Wong are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong.*
- *X. Xiao is with the Department of Integrative Biology and Physiology, University of California Los Angeles, 611 Charles E. Young Drive, Los Angeles, CA 90095-1570.*

Manuscript received 1 Jan. 2013; revised 4 May 2013; accepted 6 May 2013; published online 29 May 2013.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2013-01-0002.
Digital Object Identifier no. 10.1109/TCBB.2013.60.

microarray (PBM) [4], provide TF-TFBS binding sequence data for full-length TFs and TF domains. Note that ChIP-seq and PBM only provide high resolution on one side for the TFBSs, and there are no sequence-level experiments to directly dig out TF-TFBS binding cores. Diverse TF-TFBS binding sequences curated from peer-reviewed publications and experiments are collected in databases [17], [18]. Among the most representative ones, TRANSFAC [17] contains comprehensive published eukaryotic TF full sequences, their bound TFBSs and regulated genes. TFBSs experimentally bound by the same TFs are compiled into motifs, in the form of position weight matrices (PWMs) and consensus strings. Although sequence-level data do not directly provide the detailed information of binding cores, they provide the binding relationships, and serve as the most widely available information for discovering binding patterns (the so-called motifs) as well as the corresponding motif occurrences (the so-called instances). Regarding the data in recent work [19], there are 7,664 TF and 26,786 TFBS sequences from TRANSFAC while there are only 1,290 3D protein-DNA structures for verification. Nevertheless, PDB data serve as the most accurate verification sources to verify predicted (associated) TF-TFBS patterns as binding cores [19]. The PBM data available at UniProbe [4] can also be used to verify associated patterns on the TFBS side (but not the TF side) for their in vitro binding affinities.

The recent associated TF-TFBS pattern discovery [10], [19], [20], [21] performs two-sided motif discovery and is able to identify binding cores without any 3D structure or binding domain information. Different from structure-based methods limited by available PDB data [13] and numerous one-sided motif discovery methods designed for only TFs or TFBSs [1], the recent associated TF-TFBS pattern discovery methods work on binding sequences and employ association rule mining [10], [21] or link both TF motifs and TFBS annotations [19] to discover two-sided patterns. The resultant patterns, discovered without requiring PDB training data [8], [9], [22], familial specific information or binding domain knowledge [7], [23], [24], turn out to be verifiable binding cores [19] and positively reflect familial bindings [5].

However, none of the existing related methods [5], [10], [19], [20], [21] has proposed a formal unified model to quantitatively evaluate an associated pattern as a whole. While these methods either employ association rule mining measures such as support and confidence [25] or take advantage of existing annotated TFBS motifs in TRANSFAC [5], [19], the associated patterns cannot be scored or ranked against each other directly and quantitatively. Therefore, unified scores are desirable to shortlist and prioritize the top associated patterns for further analysis or experiment verification. Unified scores are especially useful for discovering novel associated TF-TFBS patterns in scenarios where annotations are noisy or not available, as more and more high-throughput data are being generated [4], [14]. In the long run, unified scores can serve as the basis for advanced associated pattern modeling to better understand regulatory and disease mechanisms as compared to one-sided modeling.

In this paper, we propose two unified scores (namely "sum" and "normalized" scores to be introduced) to evaluate associated TF-TFBS patterns and develop an effective pipeline to link up the TF and TFBS motif instances to form and score associated patterns. Following the background of related methods in Section 2, the methodology is detailed in Section 3. Evaluation results reported in Section 4 show that the top high-unified scores have excellent match with the top high-verification performance on existing PDB 3D structures and in extended evaluation. We discuss and conclude the paper in Section 5.

2 Background

In this section, existing methods related to associated TF-TFBS pattern discovery are briefly introduced, followed by the motivations of unified scores to model associated patterns.

2.1 Binding Residue Prediction Using 3D Structures

Available 3D structures enable training-based methods to predict binding residues on the TF (protein) side [8], [9]. They are mainly supervised methods using existing PDB 3D structures as direct training samples or employ properties derived from 3D structures such as secondary structures and solvent accessibility [22], [26]. Three-dimensional structure-based methods also suffer from the limited amount of available 3D structures to derive sophisticated features [26] and potential overfitting problems. Novel discoveries of binding cores can be prohibited for cases with no 3D structures available, such as the WRKY binding domains [27], [28]. Moreover, they are considered one sided rather than both sided as they usually only predict whether individual (or di-nucleotide-specific [26]) TF residues bind or not rather than their associated binding patterns.

2.2 One-Sided Motif Discovery Methods

Motif discovery [29] aims at finding unknown patterns (de novo motifs) and identify the corresponding motif occurrences (instances) from a set of protein or DNA sequences (one sided). Conservation and overrepresentation are the two main properties to exploit such that the discovered patterns and instances match real TF binding domains or bound TFBSs. While TF motif discovery is quite mature to discover TF binding domains with lengths of tens of residues [30], [31], TFBS motif discovery is still active and challenging [32]. There are hundreds of TFBS motif discovery methods, ranging from suffix-tree based, deterministic ones [33], [34], to artificial intelligence based, stochastic ones [30], [35]. A number of comprehensive surveys can be found [1], [29], [36]. Motifs are represented as consensus strings or PWMs of the residue distributions [29], and great challenges exist in both modeling (scores) and optimization (search strategies) to identify biologically meaningful motifs and regulatory elements. A major limitation for existing one-sided motif discovery is the lack of linkage between TFs and TFBSs to reveal information of the two-sided TF-TFBS binding cores. Nevertheless, the abundant motif discovery methods provide a wide spectrum of one-sided scores to rank and shortlist potential motifs, such as information content [37], maximal likelihood

[30], KL-divergence [38], and log odds [34]. Bayesian scores (see methods) not only are very close to most of the representative scores [38], but also balance conservation and overrepresentation well in real and noisy case studies [39], [40]. There are also consistent advances on motif representation and modeling [41], [42], [43], [44] potentially applicable to both TF and TFBS motifs. Besides one-sided motif discovery, it is intuitive to consider linking TFs and TFBSs in a both-sided manner to get stronger motif signals and better understand protein-DNA interaction patterns.

2.3 Associated TF-TFBS Pattern Discovery

Associated TF-TFBS pattern discovery distinguishes itself from existing structure-based methods and one-sided motif discovery via performing two-sided motif discovery on TF-TFBS binding sequences without using structure or domain information. With binding sequence data widely available in databases such as TRANSFAC, short and highly conserved TF and TFBS patterns on both sides can be better exploited than on only one side to reveal intriguing binding mechanisms [45]. Recently, emerging associated TF-TFBS pattern discovery methods [10], [19], [20], [21] discover very encouraging patterns that are verifiable binding cores according to PDB 3D structures as well as testable candidates supported by other evidence. Note the great challenge is that very short associated patterns (and their instances) with widths of just several residues are to be predicted using only binding sequences of full-length TFs and TFBSs, without using even domain or familial knowledge. The patterns (and their instances) are then evaluated against short binding cores extracted independently from high-resolution 3D structures that require years of efforts to determine experimentally.

The current methods include association rule mining and semi-two-sided pattern discovery with existing TFBS annotations. Association rule mining techniques [25] were first applied on TRANSFAC [10] to discover exact TF-TFBS patterns, and later more efficient data structures were developed for both exact [20] and approximate [21] cases. Association rule mining measures such as support and confidence were used to control the resultant pattern sets but no rankings or individual quantitative scores could be given. An approximate associated pattern discovery method [19] was also developed, which took advantage of existing TFBS motif annotations in TRANSFAC on the TFBS side, and linked the TF side for associated pattern with a customized TF core motif discovery algorithm. The simple core motif discovery algorithm has shown to be considerably better in identifying binding cores in the whole framework [19] than the other methods aiming at weakly conserved and TF domain-size motifs [30], [31]. The associated patterns, discovered at the sequence level without training on any 3D structures, are shown to be highly predictive and verifiable with binding cores. Therefore, they provide better insights into core protein-DNA interactions and reveal novel TF-TFBS binding rules to guide potential experiments. Besides many other potential applications, associated TF-TFBS pattern discovery has enabled generic binding subtype analysis [5] to understand regulatory mechanisms in greater detail, complementing existing one-sided studies [46], [47] with potential applications of mechanistic and disease studies related to specific bindings.

2.4 Motivations

Despite the novelty and success, none of the associated TF-TFBS pattern discovery studies has proposed any unified scores to model an associated pattern as a whole quantitatively. They either employ multiple association mining measures that collectively cannot be ranked, or take advantage of existing TFBS motifs available from TRANSFAC [19] without two-sided unified scores.

The general measures (e.g., support and confidence) in association rule ming do not capture the biological properties of motifs directly. Different support and confidence thresholds in combination can generate different numbers of patterns that, however, cannot be ranked quantitatively against each other. While the best achievable verification performance against binding cores is shown to be promising [21], how to shortlist the best output patterns in practice (before they are evaluated against the ground truth) requires more advanced and domain-specific modeling, which is still in progress for the association rule mining-based methods as shown later in the comparisons.

On the other hand, while TRANSFAC TFBS motif annotations are ready to be used [17], they do not possess scores that can be intuitively combined with the TF core motif scores to evaluate the final associated two-sided patterns. In the previous work [19], as only the annotated TFBS consensuses (or even PWMs) were used on the DNA side, the evaluation of the predicted associated patterns against the binding cores was loose. A predicted pattern was considered correct as long as its TF side motif instances were matched and a fixed width portion of the TFBS side pattern was approximately matched the binding cores, but there was no information to evaluate the actual associated TF-TFBS instances (occurrences). In this work, more detailed and stringent evaluation will be performed on all the paired TF-TFBS motif instances that can be obtained with the to-be-proposed unified scores. As a result, the instance-level verification in this paper is a much more accurate and stringent evaluation standard than the previous semipattern level verification. Moreover, as more and more raw binding data are being generated from high-throughput experiments, novel associated patterns may not be discovered only relying on noisy annotations done one sided in the past.

Unified scores for modeling associated patterns quantitatively are essential from several aspects. Because shortlisting predicted patterns for further investigation and experiments is necessary due to limited time and resources, unified scores for both-sided associated pattern modeling are desirable to evaluate and prioritize top candidates as the most testable binding cores. The scores are the central part for modeling and discovering novel associated patterns effectively as there is increasing need to analyze data beyond using existing one-sided annotations. Unified scores will also provide the basis for more advanced direct TF-TFBS modeling as simple cross linking a small number of top TF and TFBS patterns may not generate the optimal associated patterns. As a result, we are highly motivated to propose effective unified scores for associated TF-TFBS patterns in an effective pipeline to discover associated patterns. To validate the proposed unified scores, stringent instance-level verification evaluation will be performed on comprehensive ground truth to show the effectiveness of the scores in ranking predicted patterns to match real binding cores. The detailed methodology is presented in the following section.

3 Materials and Methods

In this section, we present the methodology of modeling associated TF-TFBS pattern discovery with unified scores, followed by the details of each component.

3.1 TF and TFBS Data Sets

Similar to our previous work [10], [19], we employ TRANSFAC Professional 2009.4 [17] for our experiments. There were 7,664 TFs and 26,786 corresponding bound TFBSs in total after entries without sequences were discarded. To retain high-quality data, only TFBS sequences no shorter than 8 with TRANSFAC quality levels 1-3 (smaller the better) were adopted. TFs with fewer than five TFBS sequences were discarded. After preprocessing, we have one TF data set with 607 full-length TF sequences (average length 488) for the whole TRANSFAC. Each TF sequence can be uniquely identified by its accession ID, for example, T00017, referred to as a TF entry. Each TF entry corresponds to a TFBS data set containing all the TFBS sequences it binds, which are labeled by their IDs, for example, R00207, R03135, ..., and the data set is named after the TF entry ID, as shown in the upper left of Fig. 1. Note the TFBS data sets contain the raw TFBS regions extracted from experiments, which are usually longer than the final TFBS motifs and can be over one hundred bp long. The TF data set and all the 607 TFBS data sets are the input to our method, and except the binding sequence information, no extra TF family, domain, or structure information is used. The processed input data are available in our supplementary website, which is available at http://www.cse.cuhk.edu.hk/~tmchan/patternscores/.

The overall method is shown in Fig. 1. First, extended core motif discovery based on our previous work [19] is applied on the TF data set and all the 607 TFBS Data sets, respectively (Fig. 1, Part 1). Then, all the TF and TFBS motifs are linked to form the associated patterns and calculate the unified scores (Fig. 1, Part 2). Because of additivity of the proposed unified scores, the associated pattern can be efficiently scored from the TF and TFBS scores. All associated patterns are then ranked and the top N patterns are shortlisted as the output, where N is user specified (Fig. 1, Part 3). The components are presented as follows:

3.2 TF and TFBS Core Motif Discovery

The effective customized core motif discovery algorithm of our previous work [19] is employed and extended for both TFs and TFBSs in the pipeline, as illustrated in Fig. 1, Part 1. Different from the existing one-sided motif discovery methods, the customized algorithm was developed to aim at very short and highly conserved patterns likely to be binding cores, to minimize the scattering of instance errors, and to consider the hydrophobicity properties of the TF motifs. As a result, it has showed significantly better verification performance in associated TF-TFBS pattern discovery on the TF side than the existing methods [30], [31], which on the opposite aim at long and weakly conserved domain-level motifs. Moreover, by extending our previous effective algorithm, we can better focus on the effectiveness of the proposed unified scores in modeling associated TF-TFBS patterns.

The core motif discovery algorithm is extended to work on the TFBS sequence data sets, as TRANSFAC TFBS motif annotations are no longer used. In the TFBS data sets, the raw TFBS region sequences are much shorter than those input sequences expected by a traditional TFBS motif discovery algorithm. Therefore, sophisticated methods considering long input and weak motifs [30], [40], [42], [48] may be overkills for the short and highly conserved core motifs. It is also beneficial to consistently employ the core motif discovery algorithm on the TFBS side to better focus on the unified scores in modeling. On the other hand, advanced TFBS motif discovery methods targeting for short motifs [34], [49] can be further investigated to improve and generalize the scores. For example, the Weeder released version discovers short TFBS motifs voted from several similar ones [34], and unifying the different log odds from multiple motifs is worth investigating for associated TF-TFBS patterns in future work.

The basic core motif discovery algorithm was detailed in [19]. It accepts two major parameters: the motif width W and the maximal error E allowed for any TFBS instance. For input sequences of alphabet Σ (either amino acids for TFs or nucleotides for TFBSs) and any W-width motif with a set of motif instances (the so-called answers) A, each with Hamming distance $\leq E$ from the motif, the algorithm iteratively chooses a subset A' from A to maximize the Bayesian motif score [39] Scr as follows:

$$Scr = |A'|\left(\sum_{a=1}^{W}\sum_{b\in\Sigma}\Theta_{a,b}\log\frac{\Theta_{a,b}}{\Theta_{0,b}} + \log\frac{p}{1-p} - 1\right). \qquad (1)$$

Θ is the position weight matrix of A', where $\Theta_{a,b}$ represents the frequency of residue $b \in \Sigma$ at column $a \in [1, W]$, and $\Theta_{0,b}$ is the background (i.e., input) frequency of residue b. $|S|$ is the total residue number of the data set, and $p = |A'|/|S|$ is the abundance ratio. The score Scr reflects log posterior probability of having Θ and A' with a noninformative prior and captures both conservation and overrepresentation.

The algorithm is extended as follows: Different from previous work, *ALL* similar core motifs are merged into nonredundant ones and output, each containing its instances and sequence labels as illustrated in the top right of Fig. 1. By doing so, we will not miss any potential TF or TFBS side motif candidates for the association part. To remove redundant core motifs, if any two motifs are with Hamming distance $\leq$20 percent of W or they share $\geq$80 percent instances, only the one with higher Scr will be kept. It is a natural requirement that a nontrivial motif has to have more than one instance (corresponding to minimal support $minSupport = 2$ in association rule mining). Naturally, the hydrophobicity properties check for amino acids is applicable to only TF cases, while DNA reverse complements are considered only for TFBS cases.

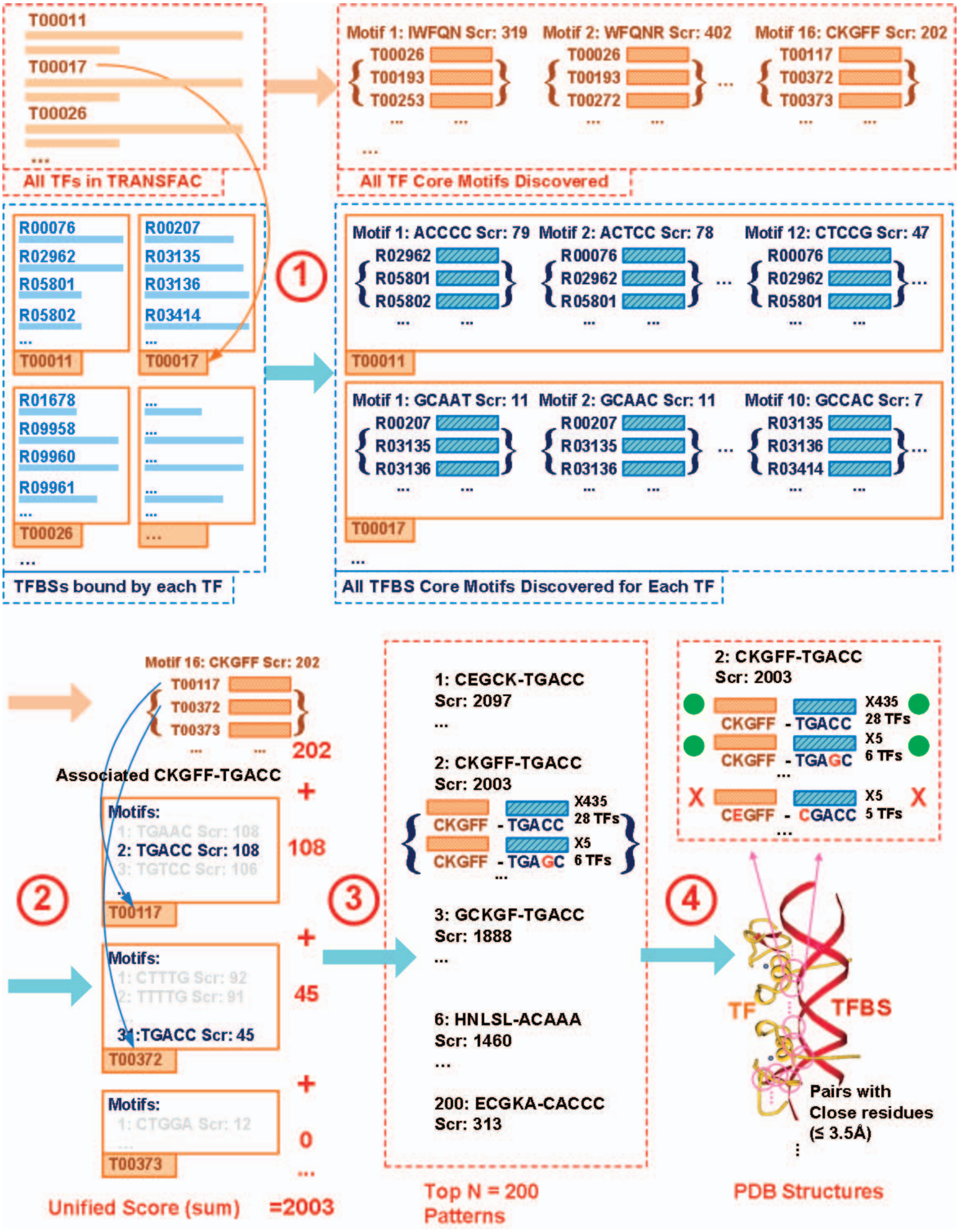


Fig. 1. Associated TF-TFBS pattern discovery illustration (motif width $W = 5$, and maximal instance error $E = 1$ for both the TF and the TFBS sides): 1. Core motif discovery on TFs and TFBSs, respectively; 2. Association of TF and TFBS core motifs based on the unified score (sum); 3. Top N associated patterns ranked and output; 4. Evaluation of verification ratios with PDB structures. The figure reflects real TRANSFAC statistics and scores (rounded), except the illustration in the evaluation. The number following X, for example, X435, means the number of actually paired TF-TFBS instances (occurrences), which is used for stringent instance-level verification. The TF number below, for example, 28 TFs, indicates the number of distinct TF entries involved.

3.3 Unified Scores for Associated Patterns

The proposed unified (two-sides) scores are calculated when the corresponding TF and TFBS motifs are associated together, as illustrated in Fig. 1, Part 2. For each TF core motif T output by the customized algorithm in Part 1, its Bayesian motif score is $Scr(T)$, for example, Motif 16: CKGFF Scr: 202 (rounded for simplicity). Each motif instance T_i of T corresponds to a TF sequence (entry), which has the corresponding TFBS data set, for example, T00117. For the TFBS data set, TFBS core motifs $\{C_i\}$ have been scored and output, for example, $\{TGAAC, TGACC, TGTCC \ldots\}$ for T00117. For every possible associated pattern T-C, for example, CKGFF-TGACC, we calculate its unified score and obtain their corresponding TF-TFBS instances (occurrences). We propose two unified scores based on combining the TF and TFBS motif Bayesian scores in an additive manner. First, the "sum" unified score is proposed and defined as the sum of the TF motif score and all the corresponding TFBS motif scores

$$sum(T\text{-}C) = Scr(T) + \sum_{T_i} Scr(C \text{ in } T_i), \qquad (2)$$

where T_i represents the TFBS data set named after the TF entry that is instance i of TF motif T. For example, the sum score for CKGFF-TGACC is $Scr(CKGFF) = 202$ plus $Scr(TGACC) = 108$ ranked second in T00117, plus $Scr(TGACC) = 45$ ranked 31st in T00372, plus 0 as TGACC is not found in T00373..., resulting in $Scr(CKGFF\text{-}TGACC) = 202 + 108 + 45 + 0 + \cdots = 2{,}003$. Two motifs may be reverse complements of each other, and only the one with the higher *Scr* is added once—they may have different scores because their log ratios against the background may be different. While $Scr(C \text{ in } T_i)$ ($= 0$ if C is not present in T_i) can be treated as independent and added, adding $Scr(T)$ may impose oversimplistic independence on the TF-TFBS dependence. Nevertheless, the proposed sum with additivity enables efficient computation and shows very promising verification results. It serves as a baseline model for more advanced scores in the future.

We also define another heuristic "normalized" unified score to balance different TFBS data set sizes as follows:

$$norm(T\text{-}C) = Scr(T)/N(T) + \sum_{T_i} Scr(C \text{ in } T_i)/N(T_i), \qquad (3)$$

where $N(T)$ means the number of sequences in the TF data set of T, i.e., $N(T) = 607$ in our experiments, and $N(T_i)$ the number of sequences in the TFBS data set of the TF entry T_i.

With the additivity of the proposed unified scores, the TF and TFBS core motif discovery parts can be done independently, and effectively linked afterward to calculate the unified scores. While nonadditive scores are also possible choices, current additive scores ensure efficient computation for the top-scored associated patterns from the top one-sided patterns without exhaustive enumerations, and show encouraging verification results as presented later. We can easily make tradeoffs about how many TF or TFBS motifs to consider in association. In our setting, we consider up to the top 10,000 TF motifs and for each T_i data set, the top 10 TFBS motifs for efficiency, because each TFBS data set corresponds to one particular TF and there are not likely many top TFBS motifs. Although the contribution of $Scr(TGACC) = 45$ ranked 31st in T00372 would be missed, candidates ranked low generally have insignificant effects on rankings, for example, the actual sum score 2,003, calculated without T00372, is the second highest.

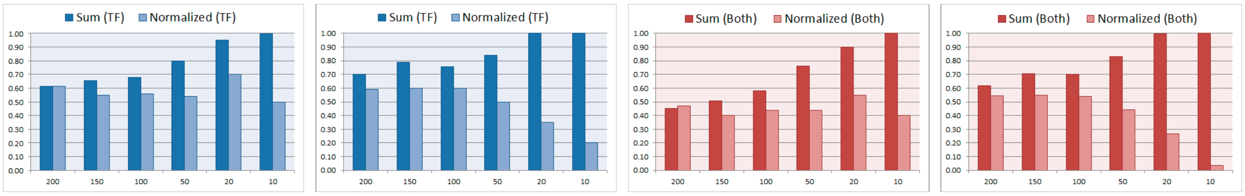
The associated patterns with their paired TF-TFBS instances are then ranked and output, as illustrated in Fig. 1 Part 3. Different from our previous work [19] that only kept the TFBS consensuses, our current pipeline not only keeps the associated patterns but also retrieves the actually paired up (associated) TF-TFBS instances according to their binding relationship in the experiment data. As a result, more detailed and stringent verification is enabled to evaluate the unified scores. In particular, all the TF-TFBS binding dependency at the instance level is maintained, and the number of actually paired TF-TFBS instances (occurrences) as well as the number of distinct TF entries involved are stored for each pattern. For example, in Fig. 1 Part 3, pair CKGFF-TGACC has X435 instances involving 28 (distinct) TF entries.

To remove noise and shortlist the top patterns, various control settings are employed. To remove potential sample noise, we introduce a min count threshold M. If an associated pattern has none of its instance pairs with a TF entry count $\geq M$, the pattern is discarded. We evaluated $M = 5, 7$ in our parameter analysis. The top N ($= 200$ in the example) patterns to output can be set considering various resource limits and priorities. For each unique and nonredundant TF core motif, we can choose the only one $K = 1$ or multiple (e.g., $K = 5$) of the top scored TFBS core motif (s) to be associated. By setting $K = 1$, we have a very stringent selection criterion keeping only one top TF-unique associated pattern; while setting $K = 5$, we have more associated pattern candidates for the same unique TF core motif. If different K settings show high verification performance, then that means the unified scores reflect the true binding cores accurately and consistently.

3.4 Evaluation with Binding Cores from PDB Structures

The predicted associated TF-TFBS patterns are evaluated with the binding cores extracted independently from PDB structures in a stringent instance-level manner. Since in our previous works, the associated patterns, discovered from sequences only, have shown to match well with familial and domain-level information [5], [10], [19], we directly evaluate them using the most precise criteria of matching the instances with binding cores, which are the tiny critical interaction fractions (interfaces) extracted from the high-resolution 3D structures. *The PDB binding core verification used here is independent and the most stringent ever for evaluating associated TF-TFBS pattern discovery* [10], [19], [21].

To evaluate an associated patterns with respect to their TF-TFBS instance pairs, binding cores (protein-DNA sequence pairs surrounding their bonding residues) from the 3D structures in PDB were extracted as the verification data following our previous work [19]. Forty thousand two hundred and twenty-two pairs were extracted from the 1,290 PDB protein-DNA complex structures. Each pair consists of two associated protein and DNA substrings (both with widths 9 in our experiments to tolerate shifted cores discovered) where the closest atom pair of the center residues is within 3.5 Å [8], [9]. There are two verification measures: one on the TF side (one sided) and the other on both TF-TFBS sides (two sided). For instance pair t-c with instance count x, for example, CKGFF-TGACC with count 435, if t on the TF side is found to be contained in certain interacting protein-DNA pairs, the pair is verified on the TF



(a) R_{TF} for $W5E0$ (b) R_{TF} for $W5E1$ (c) R_{both} for $W5E0$ (d) R_{both} for $W5E1$

Fig. 2. Comparison of verification ratios between sum and normalized unified scores ($W = 5$, $K = 5$). Vertical axis indicates the verification ratio and horizontal axis indicates N.

side. Given a verified t, its paired c on the TFBS side is further check if it appears in the corresponding protein-DNA pairs within maximal error E. If so, the whole pair t-c is verified on both sides with x instances. The count of all instance pairs verified on the TF side (both sides) over all instance pair count represent the TF verification ratio (percentage) R_{TF} (TF-TFBS both-side verification ratio R_{both}, *the most stringent measure*). For example, pair CKGFF-TGACC is verified on both sides (including the TF side), illustrated by green circles on the left and right, respectively. CKGFF-TGAGC with count 5 is also verified on both sides approximately ($E = 1$). CEGFF is not verified so the whole pair with count 5 is not verified, illustrated by red crosses on both sides. If there are only these three unique instance pairs, $R_{TF} = R_{both} = (435 + 5)/(435 + 5 + 5) = 0.99$. For N output associated patterns, we can get the average R_{TF} and R_{both} for evaluation. PDB structures are invaluable but not complete verification sources to evaluate the discovered patterns. We will introduce extended verification in the experimental results.

Note that both percentages R_{TF} and R_{both} here are instance-level measures, more stringent and more precise of prediction performance than pattern-level or semipattern level measures. In previous work [19], R_{TF} was on the instance level, but the relaxed semipattern level $R_{TF-TFBS}$ did not have nor evaluate the TFBS instance information for paired TF-TFBS instances. In particular, an associated pattern was considered verified on both sides more easily, as long as the TFBS side consensus ($\geq W$) taken from TRANSFAC could be partially verified (any substring in width W) after the TF side had been verified. Degenerate IUPAC ambiguity residues frequently happened and the old verification criteria were loose. In this work, all TFBS instances belonging to an associated pattern need to be examined exactly in width W, and any unmatched paired instances would strictly decrease R_{both}.

4 Results and Analysis

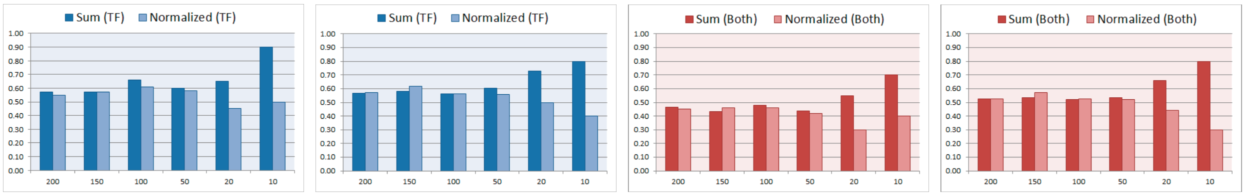
In this section, we introduce the experiment settings, report the experimental results and evaluate the performance with extended and PBM data verifications introduced.

4.1 Experiment Settings

We experimented widths $W = 5, 6$ and maximal errors $E = 0, 1$ for both TF and TFBS core motifs. The same settings were employed for both TF and TFBS sides because they were considered equally important and conserved as the potential binding cores. While we employed the intuitive settings consistent to our previous works to better focus on the unified scores in this study, different W and E settings for TFs and TFBSs can be explored in the future work. The short form, $W5E0$, for example, represents settings with $W = 5$, $E = 0$. The sum and normalized unified scores were compared and analyzed. For each nonredundant TF core motif, it could be associated with multiple ($K = 5$) or the single ($K = 1$) top TFBS core motifs in the experimental results. Different top N scored associated patterns were output and evaluated, ranging from 200 to 10. Min count thresholds $M = 5, 7$ were examined, which control that an output associated pattern must contain instance pairs with TF entry count $\geq M$. More results of N up to 500 are available in the supplementary data, available online.

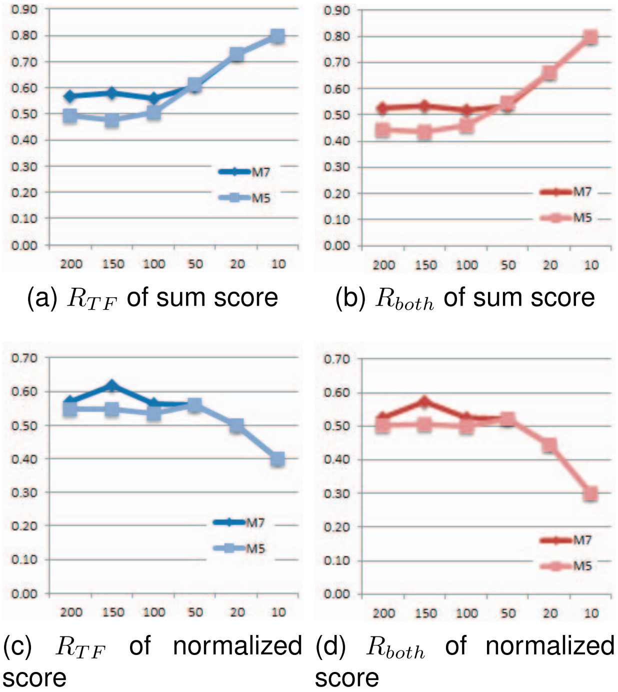
4.2 PDB Structure Verification Results

The verification performance comparisons of the "sum" and "normalized" unified scores are shown in Figs. 2 and 3 for the settings of width $W = 5$ and the min count threshold $M = 7$. The results of $K = 5$ are shown in Fig. 2. Discovered at the sequence level based purely on binding relationship without any structure, domain, or familial knowledge, the top $N \leq 200$ scored associated patterns show high verification ratios to match binding cores (interacting protein-DNA pairs) extracted from 3D structures. Note that for the stringent setting $W5E0$, TF-TFBS instance pairs of an associated pattern have to match exactly with certain



(a) R_{TF} for $W5E0$ (b) R_{TF} for $W5E1$ (c) R_{both} for $W5E0$ (d) R_{both} for $W5E1$

Fig. 3. Comparison of verification ratios between sum and normalized unified scores ($W = 5$, $K = 1$). Vertical axis indicates the verification ratio and horizontal axis indicates N.



(a) R_{TF} of sum score (b) R_{both} of sum score

(c) R_{TF} of normalized score (d) R_{both} of normalized score

Fig. 4. Examination on $M = 5, 7$. Vertical axis indicates the verification ratio (R_{TF} or R_{both}) and horizontal axis indicates N.

binding cores on both sides to be verified. For the top N results of $W5E0$ except some small variations possibly due to noises, higher unified score sums, in general, have higher average verification ratios: In Fig. 2a, $R_{TF} = 68\%$ for the top 100 patterns, 80 percent for the top 50, and 100 percent for the top 10; in Fig. 2c, $R_{both} = 58\%$ for the top 100, 76 percent for the top 50, and 100 percent for the top 10. For $W5E1$ with relaxed approximate verification on the TFBS side (TF instances still have to be exact matches), more variations are shown in the trend. The top sum score still has generally excellent correlation with the top verification ratios, for example, in Fig. 2d $R_{both} = 62\%$ for the top 200 patterns, 83 percent for the top 50, and 100 percent for the top 10. For results of $K = 1$ shown in Fig. 3, the criterion is much more stringent as only one single TFBS core motif (i.e., $K = 1$) can be selected and associated with a unique TF core motif for the highest unified score. The verification ratios decrease and more variations are shown for different N values. Nevertheless, the trend of higher sum unified scores corresponding to higher verification ratios still holds in general, where the top 10 sum scored patterns again have the best average R_{TF} and R_{both} in both $W5E0$ and $W5E1$ settings. While high normalized unified scores do not correlate with the PDB verification ratios as well as the sum scores, we will investigate into this in the extended verification section.

While R_{both} can be considered as the true-positive rate or the precision, it is not trivial to obtain the accurate false-positive rate (FPR) as the PDB 3D structures are not complete. Nevertheless, even using a conservative standard to consider 1-R_{both} as the upper bound FPR before more evidence is introduced (see Section 4.4), the verification performance for the sum score is still good with FPR ≤ 0.1 for the top 20 results ($K = 5$ in Figs. 2c and 2d).

TABLE 1
PDB R_{both} Comparison of the Top N Scored Patterns

	$W5E0$	
Method	$N = 10$	$N = 200$
sum $K = 1$	**0.70**	**0.47**
normalized $K = 1$	0.40	0.45
Association rules [21]	0.15	0.31
Annotations-used [19]	0.49 (<1.00)*	0.31 (<0.61)*

	$W5E1$	
Method	$N = 10$	$N = 200$
sum $K = 1$	**0.80**	**0.53**
normalized $K = 1$	0.30	**0.53**
Association rules [21]	N/A	N/A
Annotations-used [19]	0.07 (<0.22)*	0.25 (<0.46)*

Estimated instance-level performance; inside the parentheses is the loose semipattern level performance before normalization by the shortest TFBS consensus widths.

The sensitivity of the control parameter to remove sample noise (min count M) is investigated here. The verification ratios for thresholds $M = 5$ (M5) and $M = 7$ (M7) are illustrated with settings $W5E0$, $K = 1$ in Fig. 4. Despite the opposite trends for the different unified scores to be analyzed next, M values only slightly affect the verification performance when $N \geq 50$. For the top $N \leq 50$ results, the verification ratios are almost the same for M5 and M7. This is intuitive as patterns with more diverse TF evidence from TRANSFAC, i.e., larger M, are more likely to have available 3D structures for verification. On the other hand, $M = 7$ is more stringent and fewer patterns are output if we need up to the $N = 500$ top patterns. Other results available in the supplementary data, available online, show similar conclusions. As we focus on the top $N \leq 200$, the experiment results are all with $M = 7$.

4.3 PDB Verification Comparison with Previous Works

To compare our current work with the existing methods in Table 1, the previous loose results were reestimated according to our most stringent criteria of verification performance. As mentioned in Sections 2.4 and 3.4, the current association rule ming [21] and annotation-based pattern discovery [19] methods did not have unified scores to rank the results, and they used loose criteria to measure only the pattern- or semipattern-level verification performance. As a result, their reported evaluation results were not directly comparable to the results in this work. To demonstrate the verification performance of the unified scores, we recompiled the previous results and estimated their verification ratios R_{both} according to certain quantitative rankings.

For the association rule mining methods [20], [21] where pattern scores were not available, we employed the extended working version of [20], [21]. The current improved version has introduced p-values (in-progress details not shown) to score and rank the patterns and demonstrated the best instance-level verification performance, not just the best achievable one. Because the working version only applies to exact patterns (i.e., $E = 0$), the results corresponding to $W5E0$ were compared. P-value thresholds were selected such that the numbers of the output top $W = 5$ associated patterns were closest to

TABLE 2
Selected Protein-DNA Patterns with Annotations and Literature Surveys

Pattern		Remark
TF	TFBS	
RKQSNRESAR (Y)	CCACGTGG (Y)	Supported with homology modeling [10]
WRKYGQK (Y)	[A/G]GTCAAA (Y)	Known WRKY binding domain and W-box [28]
LQNCWSE (N)	CACGT (N)	Not in annotated domains in Uniprot [50]

(Y)—supported by literature; (N)—no known support.

$N = 10, 20$ in Table 1. The results were comparable to some $K > 1$ settings, but we just compared them to the most stringent setting of $K = 1$ in order not to favor the results of this work.

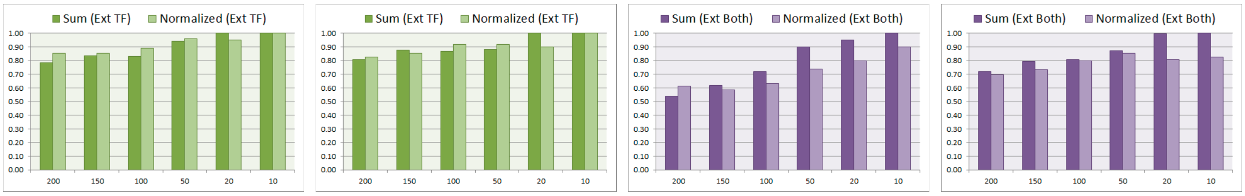
As no instance-level information was included for the annotation-based results [19], we estimated the more stringent instance-level performance. First, we removed redundancy by merging similar (with Hamming distance $\leq$20 percent of the width W) TF-side motifs. Second, we estimated their (loose) instance-level performance normalized by the minimal TFBS consensus lengths. The normalization was done because in the previous verification, the annotated TFBS-side consensus was usually longer than W and the associated pattern was considered matched as long as any W-substring of the TFBS consensus was matched with the binding cores. Finally, the top N patterns were shortlisted using the highest TF-side motif score of all the merged similar associated patterns. As a result, many similar associated patterns in [19] were merged into a single pattern under the nonredundant settings in this study, and their verification ratios were averaged and then normalized. The reported 774 patterns for $W5E0$ and the 2,559 patterns for $W5E1$ were reduced to 200 and 312 nonredundant ones respectively, comparable to $K = 1$ settings in this work.

The comparisons of TF-TFBS verification ratios R_{both} are summarized in Table 1. Without the unified scores, there is still much room for the (improved) association rule mining methods to approach the best achievable performance. With only the one-side TF scores for the TFBS annotation-based method, the top output results do not necessarily correspond to the high estimated verification performance. *The results demonstrate the importance and superiority of the unified scores to quantitatively model and evaluate the associated patterns in an overall manner*. The proposed sum score is the most effective in shortlisting the top associated patterns that are verifiable with the binding cores extracted from PDB 3D structures. Interestingly, the normalized score shows comparable performance for $N = 200$ results but much lower performance for $N = 10$. It will be investigated in the following extended verification.

4.4 Extended Verification Results and Analysis

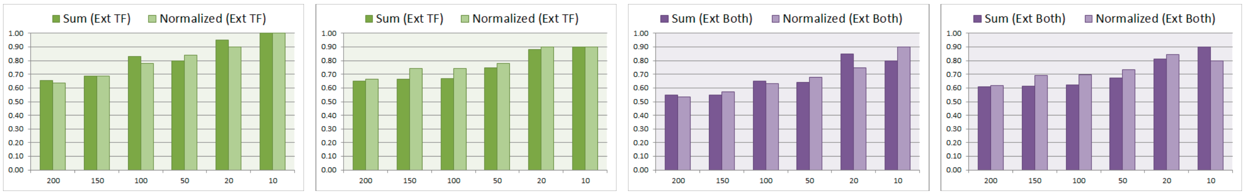
Extended verification was introduced to further analyze the top patterns not to be verified with any PDB structures. Annotations and literature surveys were employed to better evaluate the results. By grouping all 1-residue shifted patterns with the top $N = 100$ normalized scores without matching any PDB binding cores, we summarized 11 concatenated TF-TFBS patterns to be investigated with TRANSFAC and Uniprot [50] binding domain annotations as well as manual literature surveys. For a concatenated pattern, we first obtained the hosting TF information by scanning TRANSFAC, searched Uniprot, and checked if the TF-side pattern is within an annotated domain of the corresponding TF. We further checked if the specific pattern including the TFBS side is supported by literature, of not only direct interactions but also specificity critical for bindings. Part of them are listed in Table 2, with detailed information available in the supplementary data, available online. Three out of the 11 patterns do not have known support. By including the extended evidence in the verification on TF side and both sides, we obtained the average extended verification ratios R_{ExtTF} and $R_{Extboth}$ as shown in Figs. 5 ($K = 5$) and 6 ($K = 1$), respectively. Both sum and normalized unified scores show consistent and increased verification ratios. The uptrend of higher scores matching higher verification performance is better observed for normalized scores compared with Figs. 2 and 3. The upper bound FDRs are further reduced with the extended verification.

As for the best ranked patterns within the top 10 normalized scores without previous PDB verification, they are generally shorter matches to the first pattern shown in Table 2, which have been shown to match binding cores by homology modeling [10]. Another notable pattern, WRKYG-GTCAA, is ranked 50th with normalized scores and 69th with sum, respectively, in the $W5E0$, $K = 1$ settings, and ranked 96th (normalized) and 166th (sum), respectively, in $W5E1$, $K = 1$ settings. The pattern is consistently found in WRKY proteins sharing the WRKY DNA binding domains [27]. Without existing binding



(a) R_{ExtTF} for $W5E0$ (b) R_{ExtTF} for $W5E1$ (c) $R_{Extboth}$ for $W5E0$ (d) $R_{Extboth}$ for $W5E1$

Fig. 5. Comparison of extended (Ext) verification ratios between sum and normalized unified scores ($W = 5$, $K = 5$). Vertical axis indicates the verification ratio and horizontal axis indicates N.



(a) R_{ExtTF} for $W5E0$ (b) R_{ExtTF} for $W5E1$ (c) $R_{Extboth}$ for $W5E0$ (d) $R_{Extboth}$ for $W5E1$

Fig. 6. Comparison of extended (Ext) verification ratios between sum and normalized unified scores ($W = 5$, $K = 1$). Vertical axis indicates the verification ratio and horizontal axis indicates N.

structures in PDB, the invariant WRKYGQK sequence present in all WRKY domains is required for DNA binding, and in direct contact it recognizes the known W-box consensus sequence (T)TTGACY, where Y is C/T, (i.e., [A/G]GTCAA(A) in reverse complement) [28]. Therefore, the normalized score ranks WRKYG-GTCAA better than sum and both discover the real and novel binding cores within their top 100 patterns.

4.5 PBM Data Verification

As a proof of concept for the setting $W5E0$ ($K = 1$, the sum score), we employed PBM data [4] to investigate the 93 patterns (out of the 206 outputs) that were not verified even in the extended verification, i.e., $R_{Extboth} = 0$. PBM data contain in vitro DNA binding specificities for 406 nonredundant protein domains. Although TF binding core information is not readily available, TFBS 8-mers are quantitatively measured for their binding affinities. Each pattern was evaluated on the TFBS side (including reverse complements), given the TF side was matched with a particular PBM protein domain. As a result, 12 out of the 93 were verified to match the top one PBM 8-mers, and 41 verified to match within the top 10 8-mers with the highest binding affinities. Therefore, the patterns with high unified scores are likely to be real binding TFBS cores supported by PBM.

4.6 Practical Suggestions

Some suggestions are provided for practitioners to use the unified scores. In general, the sum score better correlates to and matches verification ratios and shows considerably better R_{both} than the normalized score. Both proposed scores show a high level of match with the verification performance if extra evidence is considered and are both promising as the basis models for general associated TF-TFBS patterns. In a stringent setting ($K = 1$), the normalized score sometimes outperforms the sum score in its top patterns with respect to $R_{Extboth}$. The possible reason of the discrepancy between sum and normalized scores is described as follows: The normalized score suppresses the effect of unbalanced sample sizes and favors the conservation of the patterns. While it is natural to expect that cases with more experiments done are more likely to have been investigated at the structure level, the sum score better matches verification ratios purely on PDB structures. On the other hand, the normalized score can be useful in exploring novel associated patterns without related PDB structures. In summary, the sum score is suggested to verify predicted associated pattern that have abundant binding samples and are considered to be closely related to existing evidence, for example, PDB structures. The normalized score is suggested to explore novel patterns with few related 3D structures. Besides identifying intriguing TF-TFBS binding cores demonstrated in this paper, the unified scores can be used to enrich and improve the 3D structure-based binding residue prediction [8], [26]. They can guide experiments to determine 3D binding structures [12] and also serve as a formal basis for binding (allele-specific) subtype analysis [5], [47] to decipher regulatory and disease mechanisms.

5 Discussion and Conclusion

In this paper, we have developed sum and normalized unified scores to model associated TF-TFBS patterns in general. Due to the additivity of the scores, an effective pipeline has been developed to retrieve the TF and TFBS paired instances corresponding to the core motifs discovered on each side. With the additive unified scores in association, the top associated TF-TFBS patterns can be efficiently discovered by considering the top one-sided core motifs, with no need to search low score combinations exhaustively. The scores provide accurate rankings and the method serves as a general tool for identifying binding cores and rules.

The unified sum score has shown excellent correlation and matching with high verification ratios on PDB structures. The importance of unified scores has been demonstrated in comparison with the previous methods without two-sided scores. With extended verification from annotations and thorough literature surveys, both the sum and normalized unified scores have shown consistently high verification ratios, for example, 87 and 86 percent, respectively, for the top 50 patterns under approximate settings. The top patterns discovered are confirmed to match the known WRKY binding cores that now have no available PDB complex structures. Further investigation using in vivo PBM data further confirms the effectiveness of the patterns with high unified scores.

To our knowledge, it is the first time anyone has developed unified scores to directly model associated patterns since our exploitation of binding sequences from TRANSFAC. There are some great opportunities to explore high-throughput (ChIP-seq and PBM) data beyond TRANSFAC. The associated pattern discovery methodology with the unified scores is open (processed data sets, results, and program sources are available) for improvement with more advanced modeling and efficient data structures in future work on high-throughput data such as PBM. Advanced motif models [43], [44] and discovery methods aiming at short and highly conserved motifs [34] will be explored and incorporated into the associated pattern discovery framework in the future work. The general principle of associated pattern discovery may also be applicable to other problems such as allele-specific [51] and splicing-binding [52] associations and to enhance protein-DNA binding energy modeling.

Acknowledgments

This research was supported by the Direct Grant (Project No. 2050500) from the Chinese University of Hong Kong and the GRF Grant (Project No. 310111) from the Research Grants Council of Hong Kong SAR, China.

References

[1] M. Das and H.K. Dai, "A Survey of DNA Motif Finding Algorithms," *BMC Bioinformatics,* vol. 8, no. Suppl 7, article S21, 2007.

[2] N.M. Luscombe, S.E. Austin, H.M. Berman, and J.M. Thornton, "An Overview of the Structures of Protein-DNA Complexes," *Genome Biology,* vol. 1, no. 1, 2000.

[3] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V.B. Vega, E. Wong, Y.L. Orlov, W. Zhang, J. Jiang, Y.-H. Loh, H.C. Yeo, Z.X. Yeo, V. Narang, K.R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W.-K. Sung, N.D. Clarke, C.-L. Wei, and H.-H. Ng, "Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells," *Cell,* vol. 133, no. 6, pp. 1106-1117, 2008.

[4] C. Zhu, K.J.R.P. Byers, R.P. McCord, Z. Shi, M.F. Berger, D.E. Newburger, K. Saulrieta, Z. Smith, M.V. Shah, M. Radhakrishnan, A.A. Philippakis, Y. Hu, F. De Masi, M. Pacek, A. Rolfs, T. Murthy, J. LaBaer, and M.L. Bulyk, "High-Resolution DNA-Binding Specificity Analysis of Yeast Transcription Factors," *Genome Research,* vol. 19, no. 4, pp. 556-566, Apr. 2009.

[5] T.-M. Chan, K.-S. Leung, K.-H. Lee, M.-H. Wong, T.C.-K. Lau, and S.K.W. Tsui, "Subtypes of Associated Protein-DNA (Transcription Factor-Transcription Factor Binding Site) Patterns," *Nucleic Acids Research,* vol. 40, no. 19, pp. 9392-9403, 2012.

[6] S. Mahony, P.E. Auron, and P.V. Benos, "DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies," *PLoS Computational Biology,* vol. 3, no. 3, article e61, 2007.

[7] E. Wingender, T. Schoeps, and J. Dnitz, "Tfclass: An Expandable Hierarchical Classification of Human Transcription Factors," *Nucleic Acids Research,* vol. 41, no. D1, pp. D165-D170, 2013.

[8] S. Ahmad, M.M. Gromiha, and A. Sarai, "Analysis and Prediction of DNA-Binding Proteins and Their Binding Residues Based on Composition, Sequence and Structural Information," *Bioinformatics,* vol. 20, no. 4, pp. 477-486, 2004.

[9] S. Ahmad, O. Keskin, A. Sarai, and R. Nussinov, "Protein-DNA Interactions: Structural, Thermodynamic and Clustering Patterns of Conserved Residues in DNA-Binding Proteins," *Nucleic Acids Research,* vol. 36, pp. 5922-5932, Oct. 2008.

[10] K.-S. Leung, K.-C. Wong, T.-M. Chan, M.-H. Wong, K.-H. Lee, C.-K. Lau, and S.K.W. Tsui, "Discovering Protein-DNA Binding Sequence Patterns Using Association Rule Mining," *Nucleic Acids Research,* vol. 38, pp. 6324-6337, 2010.

[11] S. Jones, P. van Heyningen, H.M. Berman, and J.M. Thornton, "Protein-DNA Interactions: A Structural Analysis," *J. Molecular Biology,* vol. 287, no. 5, pp. 877-896, 1999.

[12] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, "The Protein Data Bank," *Nucleic Acids Research,* vol. 28, no. 1, pp. 235-242, 2000.

[13] C. Winter, A. Henschel, A. Tuukkanen, and M. Schroeder, "Protein Interactions in 3D: From Interface Evolution to Drug Discovery," *J. Structural Biology,* vol. 179, no. 3, pp. 347-358, 2012.

[14] A. Valouev, D.S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R.M. Myers, and A. Sidow, "Genome-Wide Analysis of Transcription Factor Binding Sites Based on Chip-Seq Data," *Nature Methods,* vol. 5, no. 9, pp. 829-834, Sept. 2008.

[15] M. Hu, J. Yu, J.M. Taylor, A.M. Chinnaiyan, and Z.S. Qin, "On the Detection and Refinement of Transcription Factor Binding Sites Using Chip-Seq Data," *Nucleic Acids Research,* vol. 38, no. 7, pp. 2154-2167, 2010.

[16] H.S. Rhee and B.F. Pugh, "Comprehensive Genome-Wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution," *Cell,* vol. 147, no. 6, pp. 1408-1419, Dec. 2011.

[17] V. Matys, O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, and E. Wingender, "Transfac and Its Module Transcompel: Transcriptional Gene Regulation in Eukaryotes," *Nucleic Acids Research,* vol. 34, pp. 108-110, 2006.

[18] J. Bryne, E. Valen, M. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin, "JASPAR, the Open Access Database of Transcription Factor-Binding Profiles: New Content and Tools in the 2008 Update," *Nucleic Acids Research,* vol. 36, pp. D102-106, Jan. 2008.

[19] T.-M. Chan, K.-C. Wong, K.-H. Lee, M.-H. Wong, C.-K. Lau, S.K. Tsui, and K.-S. Leung, "Discovering Approximate Associated Sequence Patterns for Protein-DNA Interactions," *Bioinformatics,* vol. 27, no. 4, pp. 471-478, 2011.

[20] P.-Y. Wong, T.-M. Chan, M.H. Wong, and K.-S. Leung, "Efficient Algorithm for Mining Correlated Protein-Dna Binding Cores," *Proc. Int'l Conf. Database Systems for Advanced Applications (DASFAA '12),* pp. 470-481, 2012.

[21] P.-Y. Wong, T.-M. Chan, M.-H. Wong, and K.-S. Leung, "Predicting Approximate Protein-DNA Binding Cores Using Association Rule Mining," *Proc. IEEE 28th Int'l Conf. Data Eng. (ICDE '12),* pp. 965-976, Apr. 2012.

[22] Y. Ofran, V. Mysore, and B. Rost, "Prediction of DNA-Binding Residues from Sequence," *Bioinformatics,* vol. 23, no. 13, pp. i347-i353, 2007.

[23] S. Mahony, P.E. Auron, and P.V. Benos, "Inferring Protein-DNA Dependencies Using Motif Alignments and Mutual Information," *Bioinformatics,* vol. 23, no. 13, pp. i297-i304, 2007.

[24] S. Yang, H.K. Yalamanchili, X. Li, K.-M. Yao, P.C. Sham, M.Q. Zhang, and J. Wang, "Correlated Evolution of Transcription Factors and Their Binding Sites," *Bioinformatics,* vol. 27, no. 21, pp. 2972-2978, 2011.

[25] R. Agrawal, T. Imieliński, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *Proc. ACM Sigmod Int'l Conf. Management of Data (Sigmod '93),* pp. 207-216, 1993.

[26] M. Andrabi, K. Mizuguchi, A. Sarai, and S. Ahmad, "Prediction of Mono- and Di-Nucleotide-Specific Dna-Binding Sites in Proteins Using Neural Networks," *BMC Structural Biology,* vol. 9, no. 1, article 30, 2009.

[27] K. Yamasaki, T. Kigawa, M. Inoue, M. Tateno, T. Yamasaki, T. Yabuki, M. Aoki, E. Seki, T. Matsuda, Y. Tomo, N. Hayami, T. Terada, M. Shirouzu, A. Tanaka, M. Seki, K. Shinozaki, and S. Yokoyama, "Solution Structure of an *Arabidopsis* WRKY DNA Binding Domain," *Plant Cell,* vol. 17, no. 3, pp. 944-56, 2005.

[28] K. Yamasaki, T. Kigawa, S. Watanabe, M. Inoue, T. Yamasaki, M. Seki, K. Shinozaki, and S. Yokoyama, "Structural Basis for Sequence-Specific DNA Recognition by an *Arabidopsis* WRKY Transcription Factor," *J. Biological Chemistry,* vol. 287, no. 10, pp. 7683-7691, 2012.

[29] M. Li, B. Ma, and L. Wang, "Finding Similar Regions in Many Sequences," *J. Computer and System Sciences,* vol. 65, pp. 73-96, 2002.

[30] T.L. Bailey, "Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers," *Proc. Second Int'l Conf. Intelligent Systems for Molecular Biology,* pp. 28-36, 1994.

[31] M. Doğruel, T.A. Down, and T.J.J. Hubbard, "NestedMICA as an Ab Initio Protein Motif Discovery Tool," *BMC Bioinformatics,* vol. 9, article 19, 2008.

[32] G.K. Sandve, O. Abul, V. Walseng, and F. Drablos, "Improved Benchmarks for Computational Motif Discovery," *BMC Bioinformatics,* vol. 8, no. 1, article 193, 2007.

[33] M.F. Sagot, "Spelling Approximate Repeated or Common Motifs Using a Suffix Tree," *Proc. Third Latin Am. Symp. Theoretical Informatics (LATIN '98),* pp. 374-390, 1998.

[34] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole, "Weeder Web: Discovery of Transcription Factor Binding Sites in a Set of Sequences from Co-Regulated Genes," *Nucleic Acids Research,* vol. 32, pp. W199-W203, 2004.

[35] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, and J.C. Wooton, "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment," *Science,* vol. 262, no. 8, pp. 208-214, Oct. 1993.

[36] K.D. MacIsaac and E. Fraenkel, "Practical Strategies for Discovering Regulatory DNA Sequence Motifs," *PLoS Computational Biology,* vol. 2, no. 4, article e36, 2006.

[37] G.D. Stormo, "Computer Methods for Analyzing Sequence Recognition of Nucleic Acids," *Ann. Rev. Biophysics and Biophysical Chemistry,* vol. 17, pp. 241-263, 1988.

[38] S.T. Jensen, X.S. Liu, Q. Zhou, and J.S. Liu, "Computational Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective," *Statistical Science,* vol. 19, no. 1, pp. 188-204, 2004.

[39] S.T. Jensen and J.S. Liu, "BioOptimizer: A Bayesian Scoring Function Approach to Motif Discovery," *Bioinformatics,* vol. 20, pp. 1557-1564, 2004.

[40] T.-M. Chan, K.-S. Leung, and K.-H. Lee, "Memetic Algorithms for De Novo Motif Discovery," *IEEE Trans. Evolutionary Computation,* vol. 16, no. 5, pp. 730-748, Oct. 2012.

[41] E. Wijaya, K. Rajaraman, S.-M. Yiu, and W.-K. Sung, "Detection of Generic Spaced Motifs Using Submotif Pattern Mining," *Bioinformatics,* vol. 23, no. 12, pp. 1476-1485, 2007.

[42] T.M. Chan, G. Li, K.S. Leung, and K.H. Lee, "Discovering Multiple Realistic TFBS Motifs Based on a Generalized Model," *BMC Bioinformatics,* vol. 10, no. 1, article 321, Oct. 2009.

[43] G.D. Stormo, "Maximally Efficient Modeling of DNA Sequence Motifs at All Levels of Complexity," *Genetics,* vol. 187, no. 4, pp. 1219-1224, 2011.

[44] Y. Zhao, S. Ruan, M. Pandey, and G.D. Stormo, "Improved Models for Transcription Factor Binding Site Identification Using Nonindependent Interactions," *Genetics,* vol. 191, no. 3, pp. 781-790, July 2012.

[45] A. Sarai and H. Kono, "Protein-DNA Recognition Patterns and Predictions," *Ann. Rev. Biophysics Biomolecular Structure,* vol. 34, pp. 379-398, 2005.

[46] A.E. Kel, Y. Tikunov, N. Voss, J. Borlak, and E. Wingender, "Application of Kernel Method to Reveal Subtypes of TF Binding Motifs," *Proc. RECOMB Int'l Conf. Regulatory Genomics,* pp. 42-51, 2004.

[47] A.S.S. Bais, N. Kaminski, and P.V. Benos, "Finding Subtypes of Transcription Factor Motif Pairs with Distinct Regulatory Roles," *Nucleic Acids Research,* vol. 39, Apr. 2011.

[48] T.-M. Chan, K.-S. Leung, and K.-H. Lee, "TFBS Identification Based on Genetic Algorithm with Combined Representations and Adaptive Post-Processing," *Bioinformatics,* vol. 24, no. 3, pp. 341-349, 2008.

[49] V. Neduva and R.B. Russell, "DILIMOT: Discovery of Linear Motifs in Proteins," *Nucleic Acids Research,* vol. 34, no. Web Server issue, pp. W350-W355, 2006.

[50] UniProt Consortium, "Reorganizing the Protein Space at the Universal Protein Resource (Uniprot)," *Nucleic Acids Research,* vol. 40, no. Database issue, pp. D71-D75, Jan. 2012.

[51] G. Li, J.H. Bahn, J.-H. Lee, G. Peng, Z. Chen, S.F. Nelson, and X. Xiao, "Identification of Allele-Specific Alternative mRNA Processing via Transcriptome Sequencing," *Nucleic Acids Research,* vol. 40, no. 13, article e104, Mar. 2012.

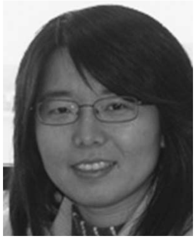
[52] Y. Wang, M. Ma, X. Xiao, and Z. Wang, "Intronic Splicing Enhancers, Cognate Splicing Factors and Context-Dependent Regulation Rules," *Nature Structural and Molecular Biology,* vol. 19, pp. 1044-1052, Sept. 2012.

Tak-Ming Chan received the BSc degree in computer science from Fudan University, China, in 2006 and the PhD degree from the Computer Science and Engineering Department, Chinese University of Hong Kong in 2010. He is currently a postdoctoral researcher in the Department of Integrative Biology and Physiology, University of California, Los Angeles. His research interests include bioinformatics and data mining.

Leung-Yau Lo received the BSc degree in risk management science from the Chinese University of Hong Kong in 2008, where he is currently working toward the PhD degree in the Department of Computer Science and Engineering under the supervision of Prof. K.S. Leung and Prof. K.H. Lee. His research interests include bioinformatics and artificial intelligence.

Ho-Yin Sze-To received the BSc degree in computer science with first class honors from the Chinese University of Hong Kong in 2011. He is currently working toward the postgraduation degree in the Department of Computer Science and Engineering, Chinese University of Hong Kong. His research interests include artificial intelligence, data mining, and machine learning as well as their applications in bioinformatics and biomedical engineering.

Kwong-Sak Leung received the BSc (Eng.) and PhD degrees from the University of London, Queen Mary College, in 1977 and 1980, respectively. He joined the Computer Science and Engineering Department, Chinese University of Hong Kong, in 1985, where he is currently a professor of computer science and engineering. His research interests are in soft computing and bioinformatics including evolutionary computation, parallel computation, probabilistic search, information fusion and data mining, fuzzy data, and knowledge engineering. He is a senior member of the IEEE.



Xinshu Xiao received the PhD degree from Harvard-MIT Division of Health Sciences and Technology in 2004. She is an assistant professor in the Department of Integrative Biology and Physiology, University of California, Los Angeles. Her research interests include genomics, bioinformatics, molecular biology, and systems biology facilitated by RNA-Seq.

Man-Hon Wong received the BSc and MPhil degrees from the Chinese University of Hong Kong in 1987 and 1989, respectively. He then went to University of California at Santa Barbara where he received the PhD degree in 1993. He joined the Chinese University of Hong Kong in August 1993 as an assistant professor. He was promoted to an associate professor in 1998. His research interests include transaction management, mobile databases, data replication, distributed systems, data mining and bioinformatics.