Heterodimeric DNA motif synthesis and validations

Ka-Chun Wong1,*, Jiecong Lin1, Xiangtao Li1, Qiuzhen Lin2, Cheng Liang3 and You-Qiang Song4

1Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong SAR, 2College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, 3School of Information Science and Engineering, Shandong Normal University, Jinan, China and 4School of Biomedical Sciences, University of Hong Kong, Pokfulam, Hong Kong SAR

Received August 14, 2018; Revised December 04, 2018; Editorial Decision December 16, 2018; Accepted December 19, 2018

ABSTRACT

Bound by transcription factors, DNA motifs (i.e. transcription factor binding sites) are prevalent and important for gene regulation in different tissues at different developmental stages of eukaryotes. Although considerable efforts have been made on elucidating monomeric DNA motif patterns, our knowledge on heterodimeric DNA motifs are still far from complete. Therefore, we propose to develop a computational approach to synthesize a heterodimeric DNA motif from two monomeric DNA motifs. The approach is sequentially divided into two components (Phases A and B). In Phase A, we propose to develop the inference models on how two DNA monomeric motifs can be oriented and overlapped with each other at nucleotide level. In Phase B, given the two monomeric DNA motifs oriented, we further propose to develop DNA-binding family-specific input-output hidden Markov models (IOHMMs) to synthesize a heterodimeric DNA motif. To validate the approach, we execute and cross-validate it with the experimentally verified 618 heterodimeric DNA motifs across 49 DNA-binding family combinations. We observe that our approach can even “rescue” the existing heterodimeric DNA motif pattern (i.e. HOXB2_EOMES) previously published on Nature. Lastly, we apply the proposed approach to infer previously uncharacterized heterodimeric motifs. Their motif instances are supported by DNase accessibility, gene ontology, protein-protein interactions, in vivo ChiP-seq peaks, and even structural data from PDB. A public web-server is built for open accessibility and scientific impact. Its address is listed as follows: http://motif.cs.cityu.edu.hk/custom/MotifKirin.

INTRODUCTION

DNA motifs are the critical components in human gene regulation; for instance, 93.2% of the DNase-accessible disease-associated SNPs were identified within DNA motifs (1). 73% of protein expressions are regulated by gene transcription in which DNA motifs play central roles (2). Most DNA motifs can be broadly classified into monomeric DNA motifs and dimeric DNA motifs. Monomeric DNA motifs are recognized and bound by the same DNA-binding proteins while dimeric DNA motifs are bound by two DNA-binding proteins in diverse orientation and spacing settings which impose a grand challenge for researchers (3).

Many high-throughput technologies have been developed to elucidate in vivo and in vitro DNA motifs such as chromatin immunoprecipitation (ChIP) followed by microarray or sequencing (i.e. ChiP-Chip, ChiP-seq and ChiP-exo) (4), microfluidic affinity analysis, protein binding microarray (PBM) (5), protein microarray assays, HT-SELEX (6), SMiLE-seq (7), and ORGANIC (8). Those high-throughput technologies have been readily adopted in the studies related to different species (especially human) nowadays; for instance, international projects (e.g. Genotype-Tissue Expression (GTEx) project, Encyclopedia of DNA Elements (ENCODE) Project, Roadmap Epigenomics Project, Dialogue on Reverse-Engineering Assessment and Methods (DREAMs), and the Functional Annotation Of Mammalian (FANTOM) genome project) have been successfully launched, leading to massive genome data accumulation at an unprecedentedly fast pace, creating opportunities for understanding DNA motifs.

To analyse those high-throughput data, scalable algorithms have been proposed and developed to identify DNA motif patterns; for instance, genome-wide DNA motif sequence patterns in human cell lines have been identified using state-of-the-arts algorithms on the ENCODE ChiP-seq data (9). Jolma et al. have also characterized different human DNA motifs from ChiP-seq and HT-SELEX data (6).
Given those precious DNA motif data, people have developed public databases and webpages for public access; for example, CIS-BP is the recently developed database which has integrated DNA motif data from multiple sources. Other databases have also been developed (namely, JASPAR, TRANSFAC, UniProbe, MotifMap, FlyTF, hPDI, ScerTF, YeTFaScO, HOCOMOCO, and TFcat) (10).

In recent years, a hidden Markov model approach has been developed for modeling adjacent nucleotide dependence and multiple motif elucidation by Wong et al. (11). Mathelier and Wasserman have also proposed a transcription factor flexible model (TFFM) to capture the position interdependence within DNA motifs in a flexible sequence length setting (12). In addition, there are other works on the simultaneous discovery of multiple DNA motifs such as MORPH, i-cisTarget, MODER, and MotifHyades (13–16). Therefore, DNA motif modeling is still a central but active problem (17).

The combinatorial interactions between multiple DNA-binding proteins have been well-noted in the past studies (18); it was assumed that multimeric DNA motif patterns are determined by the corresponding DNA-binding protein complex structures such as the cohesin (19), HNF4a (20), and HIFs (21). However, the structural determination of protein complexes are costly, labor-intensive, and time-consuming (22). Therefore, although many motifs are predicted to be bound by DNA-binding complexes in vivo, the multimeric DNA motif space has remained largely unexplored (23); for instance, 25,000 heterodimeric DNA motifs which are recognized and bound by two different types of DNA-binding proteins still have not been found in human (3). Such a situation is further complicated by the fact that some DNA motifs in the existing databases have been mixed up with heterodimeric DNA motifs without any explicit annotation from in vivo experiments (18–21,23). Such a situation is alarming and should be addressed because DNA motifs can be easily varied and related to diverse phenotypic variations across different human individuals (24).

Thanks to the recent breakthrough that Jolma et al. have proposed and performed Consecutive Affinity-Purification Systematic Evolution of Ligands by Exponential Enrichment (CAP-SELEX) on 9400 DNA motif interactions, 618 heterodimeric motifs have been successfully determined in a high-throughput manner (3). In particular, the study has revolutionized and revised our focuses from the protein structural level to the DNA sequence level, implying that the heterodimeric DNA motifs can also be determined by DNA sequence information (14). It opens a new avenue for future research since heterodimeric DNA motifs are well-known for its importance in different aspects of gene regulation (25); for instance, nucleosome displacement (26), gene expression (27), and personalized transcription factor binding repertoires (24). Therefore, we propose the first in silico approach for heterodimeric DNA motif synthesis with extensive validations for genomic insights.

This study is divided into 3 phases (i.e. Phase A, Phase B, and Phase C) as outlined in Figure 1. The objective of Phase A is to develop prediction models on the heterodimeric motif orientation and its overlapping length between two input motifs; the objective of Phase B is to develop probabilistic graphical models to synthesize heterodimeric motif patterns from two well-oriented motifs; the objective of Phase C is to apply the models developed in Phases A and B to synthesize the heterodimeric motif patterns in a DNA-binding family-specific manner.

MATERIALS AND METHODS
DNA motif data collection
We have collected the heterodimeric motif data from (3) and the related monomeric motif data from (6), resulting in 618 heterodimeric motifs and 830 monomeric motifs. For each heterodimeric motif, we retrieve its two constituent motifs using the motif names from the 830 monomeric motifs.

Input-Output hidden Markov modeling
We define the heterodimeric DNA motif synthesis problem as the input-output problem where the inputs are two motifs \{X, Y\} and the output is the heterodimeric motif \(HD\). After Phase A, \(X\) and \(Y\) are oriented and aligned as \((M_1, M_2)\) to share the same sequence length \(N\) by filling the gap positions with the background nucleotide occurring frequencies. Therefore, we can define \(M_1\) and \(M_2\) as \(4 \times N\) DNA motif model matrices where \(M_1[i, j]\) and \(M_2[i, j]\) denote the \(i\)-th nucleotide occurring fraction at the \(j\)-th position of \(M_1\) and \(M_2\) respectively. Our objective is to predict the \(HD\) as the \(4 \times N\) DNA motif model matrix output from \((M_1, M_2)\) in Phases B and C.

Obviously, the most naive solution is to take the average of \(M_1[i, j]\) and \(M_2[i, j]\) to estimate \(HD[i, j]\) where \(HD[i, j]\) denotes the \(i\)-th nucleotide occurring fraction at the \(j\)-th position of \(HD\). However, as we know, the problem context here is DNA heterodimeric motif pattern modeling. Therefore, we expect that the first part of \(HD\) usually comes from \(M_1\) and the second part of \(HD\) usually comes from \(M_2\). The overlapping region between \(M_1\) and \(M_2\) is the most difficult part to be inferred as illustrated from Figure 1.

We propose to exploit the sequence information and make the Markov assumptions to infer \(HD\) from \((M_1, M_2)\) using input-output hidden Markov models (IOHMMs) (28). Briefly, IOMMM is the generalized variant of generic
hidden Markov model (HMM). The difference between IOHMM and HMM is that IOHMM takes into account inputs and outputs with hidden states while HMM only considers outputs with hidden states. Therefore, IOHMMs are chosen in this study since we have to consider the contributions from the input DNA motif model matrices ($M_1$, $M_2$). The IOHMM mathematical modeling $θ_{IO} = (A, B, C, f, g)$ with hidden states $k \in \{1, 2, ..., K\}$ is defined as follows:


$$P(s_j = k) = f(s_{j-1}, M_1[:j-1], M_2[:j-1], j-1)$$

$$P(s_1 = k) = g(s_1, M_1[1], M_2[1])$$

$$\text{s.t. } \sum_{i=1}^{N} HD[i, j] = 1$$

$$\forall i \leq 4, j \leq N, k \leq K \in \mathbb{N}$$

where $s_j$ denotes the hidden state variable at the $j$-th position; $A_s$ is the regression coefficient matrix for $M_1$ at state $k$ in the set $A$; $B_s$ is the regression coefficient matrix for $M_2$ at state $k$ in the set $B$; $C_s$ is the bias coefficient matrix at state $k$ in the set $C$; the functions $f$ and $g$ denote the hidden state transition function and initial state estimation function respectively. In this study, given the data availability, we have adopted the classic logistic regression functions as $f$ and $g$ for computational efficiency (28).

For model training on a DNA-binding family-specific motif set ($\{M_1^{(i)}, M_2^{(i)}\}$) and $\{HD^{(i)}\}$, the Expectation Maximization (EM) algorithm is derived by taking partial derivatives to the expected complete data likelihood $E[\log L]$ (plus adding Lagrange multipliers to the probability sum to one constraints) with respect to parameters to zero. The underlying formula of the complete data likelihood $L$ is illustrated below:

$$L = \prod_{i=1}^{T} \prod_{j=1}^{N} P(s_j^{(i)}) \prod_{i=1}^{N} N(HD^{(i)}[i, j] | A_{s_j^{(i)}}[i, j] M_1^{(i)}[i, j] + B_{s_j^{(i)}}[i, j] M_2^{(i)}[i, j] + C_{s_j^{(i)}}[i, j], \sigma_{s_j^{(i)}}^2)$$

where $T$ is the total number of family-specific $\{M_1^{(i)}, M_2^{(i)}\}$ and $\{HD^{(i)}\}$; $N$ is the probability density function of the normal distribution; $\sigma_{s_j^{(i)}}^2$ is the in-sample mean squared error (or variance) of the $i$th nucleotide occurring fraction at the $j$th position of $\{HD^{(i)}\}$ at state $k$.

RESULTS

Phase A (predictions on heterodimeric motif orientation and overlapping length)

Given any two motifs, it is interesting to predict how the two motifs can be oriented and overlapped on each other if they form a heterodimeric motif. In the past, it was believed that the main driving factor should be the corresponding DNA-binding heterodimer protein structural dynamics (19–22). However, a recent ground-breaking study indicated the opposite case in which DNA motif sequences play an active role (3). Therefore, it motivates us to develop heterodimeric DNA motif synthesis models based on the readily available DNA motif data.

For each heterodimeric motif (3), we retrieve its two constituent monomeric motifs using the motif names (6) and build input features for prediction models to predict its heterodimeric motif orientation and overlapping length.

For each heterodimeric motif orientation between its two constituent motifs X and Y in the 5’ to 3’ direction, we have four possible orientation cases: X-Y, Y-X, y-X, and X-y where y is the reverse complement of Y. Owing to the double-helix nature of DNA motif, the x-Y and Y-x orientations have already been implicated by the y-X and X-y cases on the opposite strand in the 5’ to 3’ direction respectively where x is the reverse complement of X. Similarly, the x-y and y-x orientations have also been implicated by the Y-X and X-Y cases on the opposite strand in the 5’ to 3’ direction respectively.

To build models for predicting motif orientation cases, different types of input features are built for capturing global molecular dynamics. Specifically, we have carefully designed 70 sequence features (tabulated in Supplementary Table S1): numbers of motif columns, average column entropy difference, all nucleotide monomer occurring frequencies, all nucleotide dimer occurring frequencies, and DNA-binding-family-specific orientation statistics. Based on the designed features, we have implemented and compared different classifiers on each orientation case as a binary classification problem with the default parameter setting under 10-fold cross validations in R. The results are depicted in Figure 2 and Supplementary Figure S1. It is observed that random forest with 100 decision trees and logistic regression could be promising (e.g. AUROC > 0.8) in...
Phase B (probabilistic graphical modeling on heterodimeric motif patterns)

Given the orientation case and overlapping length settings between two motifs X and Y (either from prior knowledge or prediction models in Phase A), we can orient the two motifs sequentially and position the two motifs side by side as \((M_1, M_2)\). Examples are visualized on the first and second upper panels of Figures 4 and 6.

Given those two aligned motifs as the input, we can build models for heterodimeric motif synthesis. In particular, given the motif pattern degeneracy nature, we would like to focus on the probabilistic graphical modeling approaches \((11, 15)\). Specifically, to model the motif directionality and control the model complexity, we have selected the input-output hidden Markov models (IOHMMs) as our underlying model (elaborated in the previous section) since IOHMMs are designed for its sequential dependence modeling, noise tolerance, and input-driven outputs \((28)\).

Since different DNA-binding families have their own orientation and binding preferences \((3, 30)\), the previously collected 618 heterodimeric motifs \(\{HD^{(o)}\}\) have been annotated into 49 family combination groups \((3)\) as illustrated on the x-axis of Figure 5A. After the family grouping, we have trained IOHMMs on the family-specific \(\{(M_1^{(o)}, M_2^{(o)})\}\) and \(\{HD^{(o)}\}\) using the Expectation Maximization (EM) with the numbers of hidden states ranging from 2 to 10 using the default setting of depmixS4 package in R.

For each number of hidden states, we have performed 10 EM replicates to avoid any premature convergence, resulting in 90 IOHMMs for each family combination. Bayesian Information Criterion (BIC) values are calculated to choose the IOHMM model with the lowest BIC for each family combination. To estimate the performance, we have conducted leave-one-out cross-validations to infer the heterodimeric motif \(HD^{(o)}\) from the left-out pair \((M_1^{(o)}, M_2^{(o)})\) and compare \(HD^{(o)}\) with the original \(HD^{(o)}\) using average motif column distance \((11)\). In particular, we have implemented the max-product algorithm and N-max-product algorithm to elucidate the most probable hidden state transition paths in which we can derive the \(HD^{(o)}\) from the left-out pair \((M_1^{(o)}, M_2^{(o)})\) with linear complexity on each family-specific IOHMM \((11)\). Mathematically, given the trained IOHMM model \(\Theta_{IO} = (A, B, C, f, g)\) and the two input aligned matrices \((M_1^{(o)}, M_2^{(o)})\), we seek to implement the max-product algorithm to find the hidden state transition path \((k_1, k_2, \ldots, k_N)\) such that its joint occurring probability can be maximized:

\[
\arg\max_{(k_1, k_2, \ldots, k_N)} \prod_{j=1}^{N} P(s_j = k_j)
\]

As previously defined, it can be expanded as follows:

\[
\arg\max_{k_1} g(k_1, M_1^{(o)}[:, 1], M_2^{(o)}[:, 1])
\]

\[
\prod_{j=2}^{N} \arg\max_{k_j} f(k_j, k_{j-1}, M_1^{(o)}[:, j - 1], M_2^{(o)}[:, j - 1], j - 1)
\]

where we can see that the argument maximization and product steps can be computed from \(j = N\) to \(j = 2\) conditionally.
on $j - 1$ and back-tracing from $j = 1$ to $j = N$ in a sequential manner using the max-product algorithms (31). Examples are given in Figure 4. The overall results are illustrated in Figure 5A. Following our previous benchmark criteria (11), a cut-off distance of 0.5 is drawn as a red horizontal line ($P = 0.003$), indicating the robust performance of our IOHMM approach as most of them are well below the horizontal line.

Phase C (heterodimeric DNA motif synthesis)

In Phase C, we plan to apply the models developed in Phases A and B to synthesize the heterodimeric motif patterns across multiple DNA-binding family combinations.

To assess its feasibility, we have concatenated the prediction methods in Phase A and probabilistic pattern modeling approach in Phase B to form a pipeline for elucidating unknown heterodimeric motif patterns. The family-specific datasets previously described have also been adopted here. For each family combination, leave-one-out cross-validations have been conducted on its family-specific dataset to validate the pipeline.

Briefly, given two input motifs $X'$ and $Y'$, we adopt the approach in Phase A to train prediction models on the other motifs and use the trained models to predict the orientation case and overlapping length setting of $X'$ and $Y'$. After that, we orient and position $X'$ and $Y'$ as the first and second motifs ($M_1'$, $M_2'$). In the next step, we adopt the approach in Phase B to build family-specific IOHMMs on the other motifs within the same family combination and use the trained model to predict the heterodimeric DNA motif $HD'$ from the left-out pair ($M_1'$, $M_2'$) and compare $HD'$ with the original $HD$ using average motif column distance (11).

Similar to Phase B, we have implemented the max-product algorithm and N-max-product algorithm to elucidate the most probable hidden state transition paths in which we can derive $HD'$ from the left-out pair ($M_1'$, $M_2'$) with linear complexity on each family-specific IOHMM (11). The overall results are illustrated in Figure 5B. Similar to Phase B, a cut-off distance of 0.5 is drawn as a horizontal line. Most of the distance values are still below the horizontal line, confirming the good performance of our pipeline. Examples are given in Figure 6.

On the other hand, our approach also holds the novel potential for “rescuing” the existing heterodimeric motif ground-truth data from the original monomeric motif data; for instance, as shown in Figures 4D and 6D, the flanking sequence patterns on both sides have been missed in the original heterodimeric motif published on Nature (3) while our approach can recover those flanking sequence patterns from the monomeric motif data. In particular, the $5'$ flanking sequence pattern is indeed an AT-tract pattern which is well-known and important for DNA-binding recognition (32).

Applications

As for applications, we have fully trained the family-specific IOHMM models on the available heterodimeric
motif data and applied them to the 1915 no-signal heterodimers as outlined on the extended data figure 2 of (3) published on Nature. In particular, only 515 of them have the available monomeric motif information for matching. Therefore, we have applied the proposed approach to those 515 heterodimers using the respective DNA-binding-family-specific models. The synthesized heterodimeric motifs are publicly released for open accessibility in the following web address: http://bioinfo.cs.cityu.edu.hk/PhaseCrunOnNOSignalpairs_Results.zip. Its count histogram is visualized in Supplementary Figure S4.

To study those newly synthesized heterodimeric motifs, we have matched them to the existing HOCOMOCO motif database (i.e. DNA HOCOMOCO Human (v11 Full)) using TomTom (33). Interestingly, based on the default TomTom’s statistical significance testing, we found that only 17.5% (90/515) of those motifs can be matched to both of the constituent motifs while half of them 50.1% (258/515) can be matched to one of its constituent motifs. The remaining 32.4% (167/515) are no longer matching to its constituent motifs. It implies that those newly synthesized heterodimeric motif patterns can be interesting.

To validate those motifs in a genome-wide manner, we rely on FIMO with its default setting (34) to scan each motif on the whole human genome (hg19) and computed its motif instances’ DNase accessibility (i.e. ENCODE DNase cluster peaks) and evolutionary conservation (i.e. PhyloP100way).

Those motif instances’ DNase accessibility (i.e. ENCODE DNase cluster peaks) are summarized in Figure 7. Interestingly, we can observe that those heterodimeric DNA motif instances are more overlapped with the DNase cluster peaks than the human genomic background. It indicates that those instances can be accessible on chromatin, reflecting its genomic relevance.

The evolutionary conservation (i.e. PhyloP100way) of those genome-wide motif instances are also summarized in Supplementary Figure S5. Surprisingly, unlike the DNase accessibility, the motif instances do not show any strong evolutionary conservation signal. One possible explanation is that heterodimeric DNA motif instances are known for its DNA-binding flexibility where mismatches can be tolerated because of the elevated DNA-binding specificity of heterodimers (35). It is also consistent with the previous study that only 24% of the heterodimeric DNA motifs are enriched in genomic conservation (3).

Given its chromatin accessibility and evolutionary conservation, one may be interested in the functionalities of those newly synthesized heterodimeric motifs. Therefore, we have run GOMo with its default settings to scan those heterodimeric motifs on all human promoters, retrieving their target gene ontology (GO) terms (36). The overall results are depicted in Supplementary Figure S6. Interestingly, 96.9% (499/515) of those newly synthesized heterodimeric motifs can be associated to at least one GO term with statistical significance as reported by GOMo (36).

In particular, 28.0% (144/515) of them are associated with the DNA binding term (i.e. GO:0003677 DNA binding) on their downstream target genes; it implies that those motifs serve as the mediators for subsequent DNA-binding
interactions. Furthermore, we note that the most frequent GO term is ‘GO:0004984 olfactory receptor activity’ with 54.4% (280/515) motifs associated. In particular, we observe that 94.6% (265/280) of those motifs are bound by homeodomain-related heterodimers, suggesting their important roles in olfactory sensory systems beyond the existing knowledge limited to monomeric homeodomain proteins (37). Besides, we also observe that the organelle-related GO terms are also frequently observed; for instance, the ‘GO:0043231 intracellular membrane-bounded organelle’ term is observed among 40.6% (209/515) motifs. In particular, we note that 87.1% (182/209) of them are bound by ETS-related heterodimers; it suggests the underlying important roles of the downstream organelle activities induced by ETS-related proteins which are related to cell development and cancer progression (38). Similar observations can be made for other GO terms (e.g. GO:0004984, GO:0007600, GO:0007606, GO:0007608, and GO:0050890).

We have also checked whether the two DNA-binding proteins corresponding to each newly synthesized heterodimeric DNA motif have any reported interaction. Therefore, we have mapped them onto the STRING database (version 10; species: 9606) (39). We can observe that 26.8% (138/515) of those motifs have reported protein-protein interactions for their corresponding transcription factors. Such a statistics is meaningful as only 2.04% hits can be estimated to occur by chance, given that the interaction database has 20,457 proteins with 4,274,001 reported interactions. Therefore, our protein-protein interaction statistics are 12-fold-enriched than the random hit hypothesis (Fisher Exact Test \( P \)-value \(< 2.2 \times 10^{-16}\)). The detailed results are depicted in Supplementary Figure S7.

For the motif instances on the ELF1 ChIP-seq peaks in Supplementary Table S3, we can observe that the newly synthesized E2F3\_ELF1 and GCM2\_ELF1 motifs (as visualized in Supplementary Figure S8) occur even more frequently than the original ELF1 motif instances. It is actually surprising since the ChIP-seq experiment was designed for the single protein ELF1 but we can still observe the ubiquitous occurrences of the E2F3\_ELF1 and GCM2\_ELF1 motifs. Their values of DNase accessibility and evolutionary conservation are also much higher than the human genomic background. In addition, the STRING analysis also proves that E2F3 is associated with ELF1 as depicted in Supplementary Figure S7. It demonstrates the genomic relevance and potential of those newly synthesized heterodimeric DNA motifs.

For the motif instances on the PAX5 ChIP-seq peaks in Supplementary Table S4, we can observe that the occurring frequencies of the newly synthesized ETV5\_PAX5 and ERF\_PAX5 motifs (as visualized in Supplementary Figure S10) are surprisingly close to that of the original PAX5 motif instances. It is similar to the previous cases where the \textit{in silico} synthesized heterodimeric motifs can actually be found on the \textit{in vivo} ChIP-seq peaks, demonstrating the practical uses of the proposed approach.
In particular, we are fortunate to find that there is a heterodimeric DNA-binding complex structure available in PDB for PAX5 (PDB id: 1K78) where its binding partner is ETS1 which shares the same DNA-binding domain with ETV5. Therefore, we have retrieved the complex structure data from PDB. We are also fortunate to find the DNA-binding structure of ETV5 (PDB id: 4UNO). Therefore, we have superimposed the ETV5 structure onto the the ETS1–PAX5 complex using MatchMaker with its default setting (40). The results are depicted in Figure 8. Interestingly, we can observe that the ETV5 fits very well into the DNA-binding position of ETS1 since they belong to the same DNA-binding domain family (ETS). The un-matching regions come from the C-terminus which is far from the DNA-binding region. Therefore, we argue that our in silico synthesized ETV5-PAX5 heterodimeric motif can even be supported by the available 3D structural data, demonstrating its capacity to infer heterodimeric motifs from the existing plethora of monomeric motif data.

**DISCUSSION**

In the past years, significant efforts were devoted to individual DNA motif elucidation as limited by the past assumption that the heterodimeric DNA motifs are governed by the corresponding heterodimeric protein complex dynamics. However, such an assumption has become rather speculative in recent years.

Therefore, we proposed an approach to infer heterodimeric DNA motifs from monomeric DNA motif information. In that approach, given two monomeric DNA motifs, we have compared different classification and regression models (e.g. random forest) to predict the actual direction and spacing of the motifs. Once predicted, we have further proposed the IOHMM model to represent and synthesize a heterodimeric DNA motif.

The approach has been extensively validated on the experimentally verified 618 heterodimeric DNA motifs across 49 DNA-binding-family combinations. Its results indicate that the approach can infer the heterodimeric DNA motif patterns similar to the verified ones. In particular, it even has the potential to ‘rescue’ the existing verified ones for genomic insights.

For the applications, the proposed approach has been applied to elucidate the previously uncharacterized 515 heterodimeric DNA motifs. Their resultant motif instances on the human genome are well-supported by DNase accessibility, gene ontology terms, protein-protein interactions, in vivo ChIP-seq peaks, and even molecular structure data.

In the future, we envision that the proposed approach can be applied to other motif data for heterodimeric DNA motif pattern synthesis, enabling numerous downstream studies. On the other hand, it will be interesting to extend it for hetero-multimeric motif synthesis subject to data availability. The growth in the combinatorial space of multiple motif pattern matching can be a computational challenge.

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

**ACKNOWLEDGEMENTS**

The authors would like to thank Arthu Jolma, Jussi Taipale’s lab, and the CIS-BP team for making their DNA motif data publicly available. The authors would also like to thank Prashant Sridhar for his English proofreading. The authors would also like to thank the three anonymous reviewers for their constructive comments. This study is substantially supported by Research Grants Council in Hong Kong.

**FUNDING**

Research Grants Council of the Hong Kong Special Administrative Region [CityU 21200816, CityU 11203217, CityU 11200218]; Titan Xp GPU from the NVIDIA Corporation. Funding for open access charge: Research Grants Council (Hong Kong).

**Conflict of interest statement.** None declared.

**REFERENCES**
