1628–1636 *Nucleic Acids Research, 2019, Vol. 47, No. 4* *Published online 27 December 2018*
doi: 10.1093/nar/gky1297

Heterodimeric DNA motif synthesis and validations

Ka-Chun Wong[1,*], Jiecong Lin[1], Xiangtao Li[1], Qiuzhen Lin[2], Cheng Liang[3] and You-Qiang Song[4]

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong SAR, [2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, [3]School of Information Science and Engineering, Shandong Normal University, Jinan, China and [4]School 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).

*To whom correspondence should be addressed. Tel: +852 34428618; Email: kc.w@cityu.edu.hk

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.

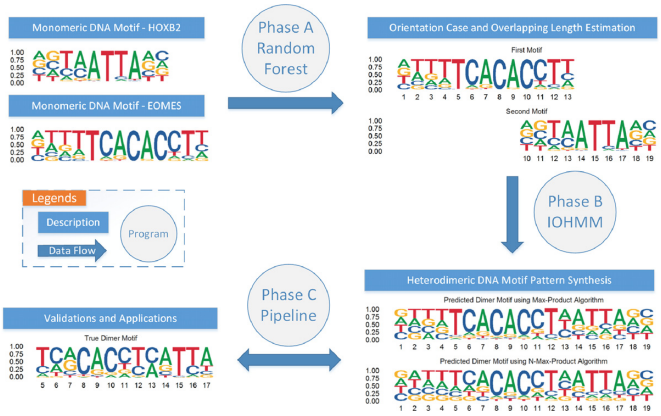

Figure 1. Overview of proposed approach. The approach is divided into three phases with distinct objectives. It is noted that the monomeric DNA motif of EOMES has been reverse-complemented for clarity.

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 ($M1$, $M2$) to share the same sequence length N by filling the gap positions with the background nucleotide occurring frequencies. Therefore, we can define $M1$ and $M2$ as $4 \times N$ DNA motif model matrices where $M1[i, j]$ and $M2[i, j]$ denote the i-th nucleotide occurring fraction at the j-th position of $M1$ and $M2$ respectively. Our objective is to predict the HD as the $4 \times N$ DNA motif model matrix output from ($M1$, $M2$) in Phases B and C.

Obviously, the most naive solution is to take the average of $M1[i, j]$ and $M2[i, j]$ to estimate $HD[i, j]$ where $HD[i, j]$ denotes the ith nucleotide occurring fraction at the jth 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 $M1$ and the second part of HD usually comes from $M2$. The overlapping region between $M1$ and $M2$ 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 ($M1$, $M2$) 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 ($M1$, $M2$). The IOHMM mathematical modeling $\Theta_{IO} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, f, g)$ with hidden states $k \in \{1, 2, ..., K\}$ is defined as follows:

$$HD[i, j] = A_{s_j}[i, j]M1[i, j] + B_{s_j}[i, j]M2[i, j] + C_{s_j}[i, j]$$

$$P(s_j = k) = f(s_j, s_{j-1}, M1[:, j-1], M2[:, j-1], j-1)$$

$$P(s_1 = k) = g(s_1, M1[:, 1], M2[:, 1])$$

$$s.t. \sum_{i=1}^{4} HD[i, j] = 1$$

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

where s_j denotes the hidden state variable at the j-th position; A_k is the regression coefficient matrix for $M1$ at state k in the set $\mathbf{A}$; B_k is the regression coefficient matrix for $M2$ at state k in the set $\mathbf{B}$; C_k is the bias coefficient matrix at state k in the set $\mathbf{C}$; 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 $\{(M1^{(t)}, M2^{(t)})\}$ and $\{HD^{(t)}\}$, 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:

$$L = \prod_{t=1}^{T}\prod_{j=1}^{N} P(s_j^{(t)}) \prod_{i=1}^{4} \mathcal{N}(HD^{(t)}[i, j] \mid A_{s_j^{(t)}}[i, j]M1^{(t)}[i, j] + B_{s_j^{(t)}}[i, j]M2^{(t)}[i, j] + C_{s_j^{(t)}}[i, j], \sigma^2_{s_j^{(t)}}[i, j])$$

where T is the total number of family-specific $\{(M1^{(t)}, M2^{(t)})\}$ and $\{HD^{(t)}\}$; $\mathcal{N}$ is the probability density function of the normal distribution; $\sigma_k^2[i, j]$ is the in-sample mean squared error (or variance) of the ith nucleotide occurring fraction at the jth position of $\{HD^{(t)}\}$ 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.

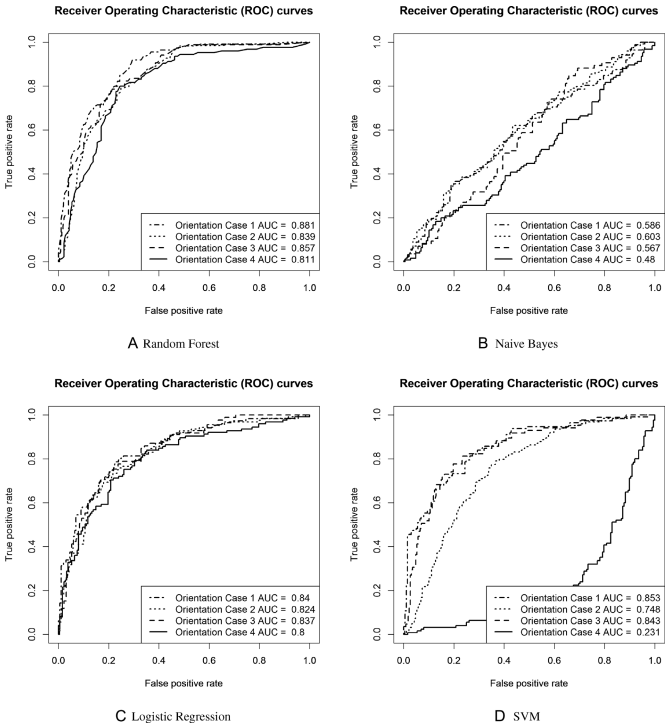

Figure 2. Receiver-operating characteristic (ROC) curves for predicting orientation cases under 10-fold cross-validations in R default.

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

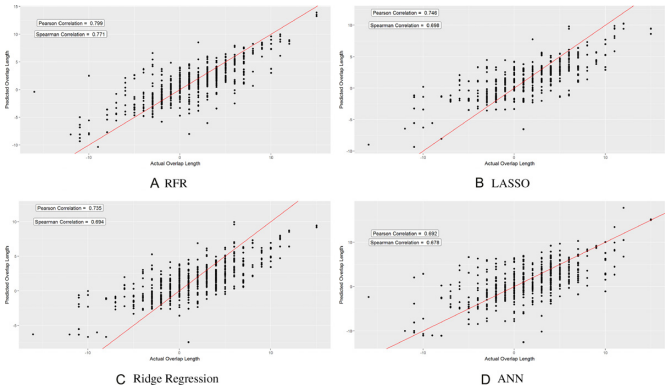

Figure 3. Scatter plots for predicting overlapping lengths under 10-fold cross-validations where RFR stands for random forest regression and ANN stands for artificial neural network. Negative values indicate non-overlapping spacing lengths. Red lines denote the ideal cases. Default parameter settings in R are adopted.

this task. To reveal mechanistic insights, we have performed random forest feature importance analysis (29) as visualized in Supplementary Figure S2. Different feature priorities are observed for different cases.

Once the orientation case has been identified, the next step is to determine how many nucleotide positions are overlapped between the first motif and the second motif in the 5′ to 3′ direction. From the machine learning perspective, we consider it as a regression problem where the inputs are the two oriented motifs while the output is the predicted overlapping length between them. To build models for predicting overlapping lengths, we would like to put emphasis on the local sequence degeneracy compatibility at the potential overlapping locations. Therefore, we have carefully designed 82 sequence features (tabulated in Supplementary Table S2): orientation case, numbers of motif columns, DNA-binding-family-specific overlapping length statistics, average motif column entropies and its difference, all nucleotide monomer occurring frequencies, all nucleotide dimer occurring frequencies, column Euclidean distances at different possible overlapping positions, overlapping location indices with the minimal column Euclidean distances, column entropies and its differences at different overlapping positions, overlapping location indices with the maximal and minimal column entropy differences, and the cumulative averages of the previous measurements whenever applicable.

Based on the designed features, we have tried and compared different regression methods to predict the overlapping length with the default parameter setting under 10-fold cross validations in R. The results are depicted in Figure 3. It is observed that the random forest regression with 100 decision trees performed better than the others (Pearson and Spearman correlations > 0.75).

In addition, we have performed random forest feature importance analysis (29) as visualized in Supplementary Figure S3. Different feature priorities are observed again. In particular, we can see that the models heavily rely on the past family-specific information for the inference. Such knowledge is consistent with the existing belief that the DNA-binding mechanisms are family-specific, motivating us to develop family-specific models in Phases B and C.

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 ($M1$, $M2$). 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 $\{HD^{(t)}\}$ 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 $\{(M1^{(t)}, M2^{(t)})\}$ and $\{HD^{(t)}\}$ 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 $(M1^{(o)}, M2^{(o)})$ 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 $(M1^{(o)}, M2^{(o)})$ with linear complexity on each family-specific IOHMM (11). Mathematically, given the trained IOHMM model $\Theta_{IO} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, f, g)$ and the two input aligned matrices $(M1^{(o)}, M2^{(o)})$, we seek to implement the max-product algorithm to find the hidden state transition path $(k_1, k_2, ..., k_N)$ such that its joint occurring probability can be maximized:

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

As previously defined, it can be expanded as follows:

$$\arg\max_{k_1} g(k_1,\ M1^{(o)}[:, 1],\ M2^{(o)}[:, 1])$$

$$\prod_{j=2}^{N} \arg\max_{k_j} f(k_j, k_{j-1},\ M1^{(o)}[:, j-1],\ M2^{(o)}[:, j-1],\ j-1)$$

where we can see that the argument maximization and product steps can be computed from $j = N$ to $j = 2$ conditionally

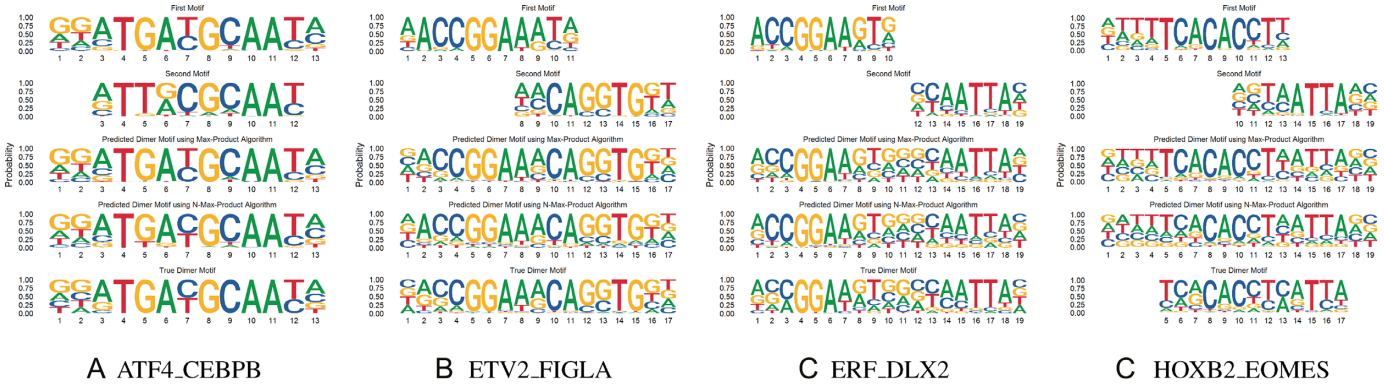

Figure 4. Heterodimeric motif synthesis cases with the true orientation case and overlap length settings under leave-one-out cross-validations.

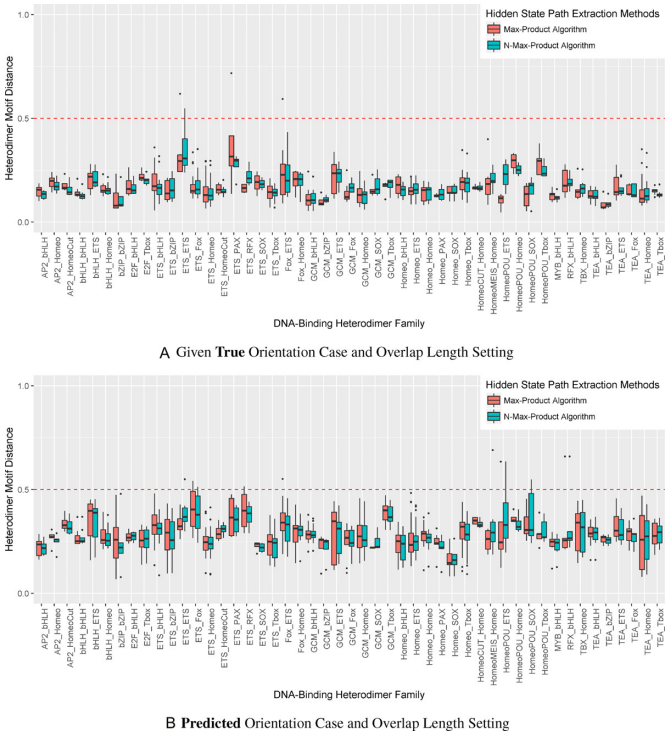

Figure 5. Family-specific average motif column distance performance distributions (measured on the vertical axes) under leave-one-out cross-validations across 49 DNA-binding family combinations (denoted on the horizontal axes) (3).

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 ($M1'$, $M2'$). 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 ($M1'$, $M2'$) 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 ($M1'$, $M2'$) 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

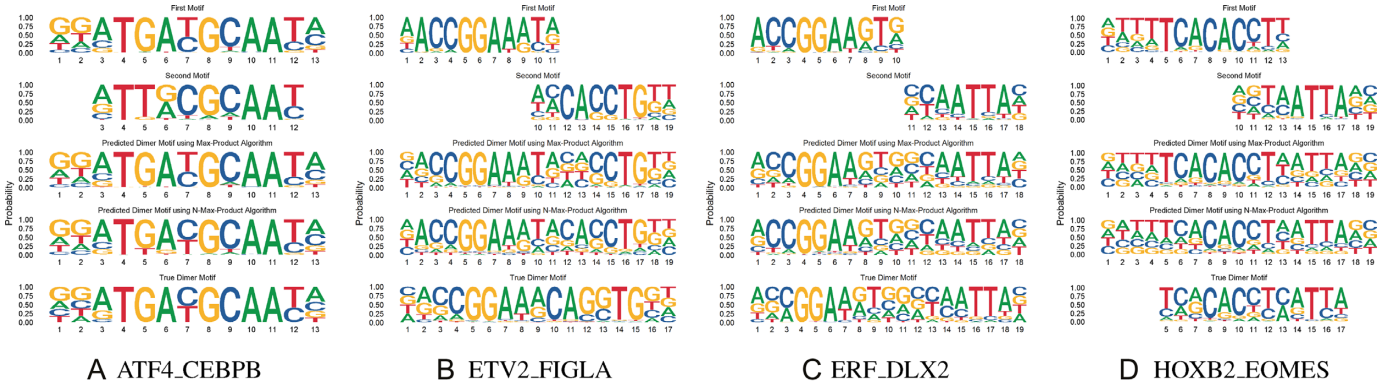

Figure 6. Heterodimeric motif synthesis cases with the predicted orientation case and overlap length settings under leave-one-out cross-validations.

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 relies 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

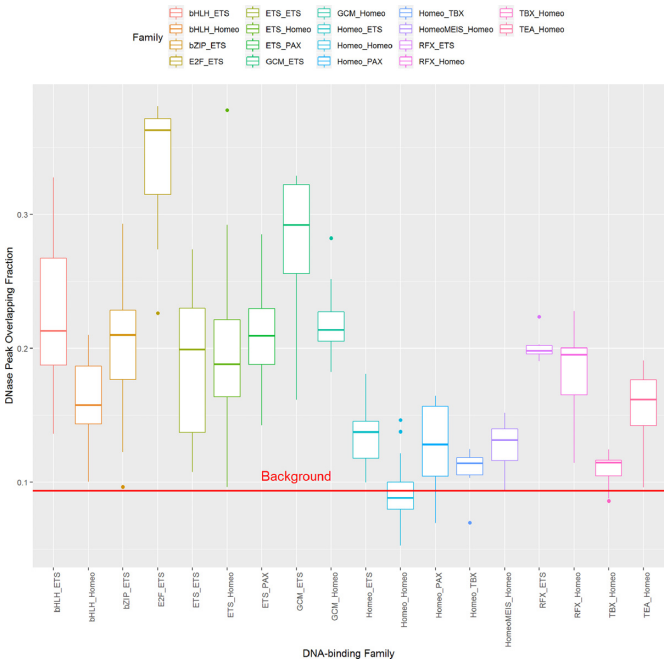

Figure 7. Boxplot on the DNase accessibility (i.e. ENCODE DNase cluster peaks) for the newly synthesized heterodimeric motifs grouped by DNA-binding family combinations. The background denotes the DNase peak coverage fraction across the whole human genome (hg19).

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

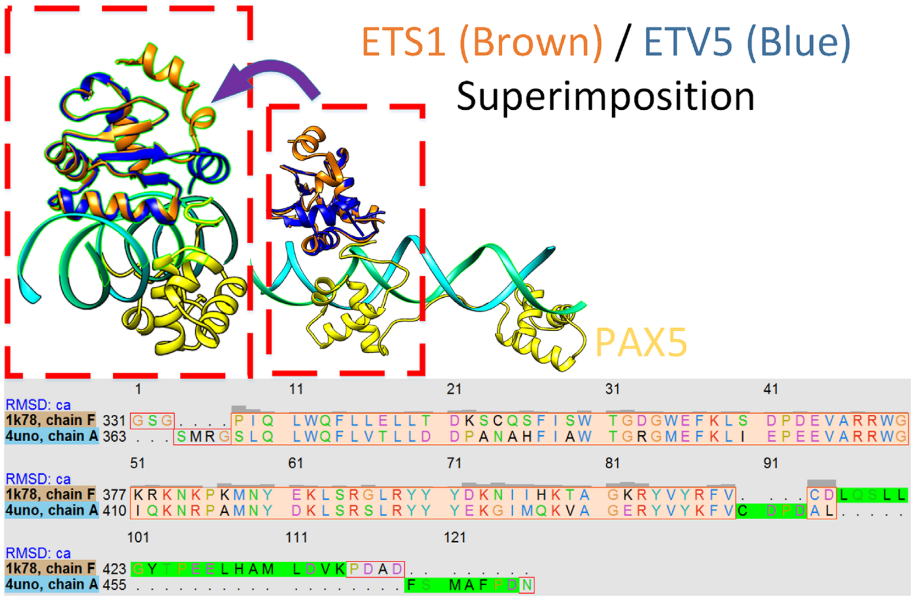

Figure 8. Superimposition of ETV5 (PDB id: 4UNO) onto the ETS1 location of the DNA-binding complex of ETS1-PAX5 (PDB id: 1K78) using MatchMaker with its default setting (40). The sequence alignment box is drawn using the "Match-Align" procedure; it indicates 53% sequence identity between ETV5 and ETS1. This whole figure is drawn using UCSF Chimera with the chain coloring option (41). The left red box is the rotated view of the right red box with light green highlights on the superimposition.

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.

To investigate those newly synthesized 515 heterodimeric motif patterns further, we have searched for its related ChIP-seq data in ENCODE. Two DNA-binding proteins (ELF1 and PAX5) have its ChIP-seq peaks available (see details in the supplementary). Therefore, we have scanned all of their related heterodimeric motif patterns in this study on those ChIP-seq peaks using FIMO with its default command-line setting. We have also computed the aforementioned DNase accessibility and evolutionary conservation to estimate the heterdimeric motif instance importance. The results are tabulated in Tables S3 and S4.

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 *in silico* synthesized heterodimeric motifs can actually be found on the *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 location of 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

1. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
2. Li,J.J. and Biggin,M.D. (2015) Gene expression. Statistics requantitates the central dogma. *Science*, **347**, 1066–1067.
3. Jolma,A., Yin,Y., Nitta,K.R., Dave,K., Popov,A., Taipale,M., Enge,M., Kivioja,T., Morgunova,E. and Taipale,J. (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, **527**, 384–384.
4. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
5. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
6. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
7. Isakova,A., Groux,R., Imbeault,M., Rainer,P., Alpern,D., Dainese,R., Ambrosini,G., Trono,D., Bucher,P. and Deplancke,B. (2017) SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat. Methods*, **14**, 316–322.
8. Kasinathan,S., Orsi,G.A., Zentner,G.E., Ahmad,K. and Henikoff,S. (2014) High-resolution mapping of transcription factor binding sites on native chromatin. *Nat. Methods*, **11**, 203–209.
9. Kheradpour,P. and Kellis,M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.*, **42**, 2976–2987.
10. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
11. Wong,K.C., Chan,T.M., Peng,C., Li,Y. and Zhang,Z. (2013) DNA motif elucidation using belief propagation. *Nucleic Acids Res.*, **41**, e153.
12. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
13. Herrmann,C., Van de Sande,B., Potier,D. and Aerts,S. (2012) i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Res.*, **40**, e114.
14. Toivonen,J., Kivioja,T., Jolma,A., Yin,Y., Taipale,J. and Ukkonen,E. (2018) Modular discovery of monomeric and dimeric transcription factor binding motifs for large data sets. *Nucleic Acids Res.*, **46**, e44.

15. Wong,K.C. (2017) MotifHyades: expectation maximization for de novo DNA motif pair discovery on paired sequences. *Bioinformatics*, **33**, 3028–3035.
16. Wong,K.C., Li,Y. and Peng,C. (2016) Identification of coupling DNA motif pairs on long-range chromatin interactions in human K562 cells. *Bioinformatics*, **32**, 321–324.
17. Wong,K.C., Li,Y., Peng,C., Moses,A.M. and Zhang,Z. (2015) Computational learning on specificity-determining residue-nucleotide interactions. *Nucleic Acids Res.*, **43**, 10180–10189.
18. Reiter,F., Wienerroither,S. and Stark,A. (2017) Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.*, **43**, 73–81.
19. Murayama,Y. and Uhlmann,F. (2014) Biochemical reconstitution of topological DNA binding by the cohesin ring. *Nature*, **505**, 367.
20. Chandra,V., Huang,P., Potluri,N., Wu,D., Kim,Y. and Rastinejad,F. (2013) Multi-Domain Integration in the Structure of the HNF4α Nuclear Receptor Complex. *Nature*, **495**, 394.
21. Wu,D., Potluri,N., Lu,J., Kim,Y. and Rastinejad,F. (2015) Structural integration in hypoxia-inducible factors. *Nature*, **524**, 303.
22. Jiang,F., Taylor,D.W., Chen,J.S., Kornfeld,J.E., Zhou,K., Thompson,A.J., Nogales,E. and Doudna,J.A. (2016) Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science*, **351**, 867–871.
23. Inukai,S., Kock,K.H. and Bulyk,M.L. (2017) Transcription factor–DNA binding: beyond binding site motifs. *Curr. Opin. Gene Dev.*, **43**, 110–119.
24. Barrera,L.A., Vedenko,A., Kurland,J.V., Rogers,J.M., Gisselbrecht,S.S., Rossin,E.J., Woodard,J., Mariani,L., Kock,K.H., Inukai,S. *et al.* (2016) Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*, **351**, 1450–1454.
25. Ravasi,T., Suzuki,H., Cannistraci,C., Katayama,S., Bajic,V.B., Tan,K., Akalin,A., Schmeier,S., Kanamori-Katayama,M., Bertin,N. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.
26. Deplancke,B., Alpern,D. and Gardeux,V. (2016) The genetics of transcription factor DNA binding variation. *Cell*, **166**, 538–554.
27. Huminiecki,Ł. and Horbańczuk,J. (2017) Can we predict gene expression by understanding proximal promoter architecture? *Trends Biotechnol.*, **35**, 530–546.
28. Bengio,Y. and Frasconi,P. (1995) An input output HMM architecture. In: *Advances in Neural Information Processing Systems*. pp. 427–434.
29. Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.
30. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
31. Frey,B.J. and Jojic,N. (2005) A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE Trans PAMI*, **27**, 1392–1416.
32. Koiwai,K., Kubota,T., Watanabe,N., Hori,K., Koiwai,O. and Masai,H. (2015) Definition of the transcription factor TdIF1 consensus-binding sequence through genomewide mapping of its binding sites. *Genes Cells*, **20**, 242–254.
33. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A. *et al.* (2017) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.
34. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
35. Chang,A.T., Liu,Y., Ayyanathan,K., Benner,C., Jiang,Y., Prokop,J.W., Paz,H., Wang,D., Li,H.R., Fu,X.D. *et al.* (2015) An evolutionarily conserved DNA architecture determines target specificity of the TWIST family bHLH transcription factors. *Genes Dev.*, **29**, 603–616.
36. Buske,F.A., Bodén,M., Bauer,D.C. and Bailey,T.L. (2010) Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics*, **26**, 860–866.
37. Hirota,J. and Mombaerts,P. (2004) The LIM-homeodomain protein Lhx2 is required for complete development of mouse olfactory sensory neurons. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 8751–8755.
38. Bose,R., Karthaus,W.R., Armenia,J., Abida,W., Iaquinta,P.J., Zhang,Z., Wongvipat,J., Wasmuth,E.V., Shah,N., Sullivan,P.S. *et al.* (2017) ERF mutations reveal a balance of ETS factors controlling prostate oncogenesis. *Nature*, **546**, 671.
39. Szklarczyk,D., Morris,J.H., Cook,H., Kuhn,M., Wyder,S., Simonovic,M., Santos,A., Doncheva,N.T., Roth,A., Bork,P. *et al.* (2016) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.*, **45**, D362–D368.
40. Meng,E.C., Pettersen,E.F., Couch,G.S., Huang,C.C. and Ferrin,T.E. (2006) Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics*, **7**, 339.
41. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.