Published online 20 September 2013

Nucleic Acids Research, 2013, Vol. 41, No. 21 ***e197***
doi:10.1093/nar/gkt831

A general approach for discriminative *de novo* motif discovery from high-throughput data

Jan Grau[1], Stefan Posch[1], Ivo Grosse[1] and Jens Keilwagen[2,3,*]

[1]Institute of Computer Science, Martin Luther University Halle–Wittenberg, D-06099 Halle, Saale, Germany, [2]Institute for Biosafety in Plant Biotechnology, Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, D-06484 Quedlinburg, Germany and [3]Department of Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D-06466 Seeland OT Gatersleben, Germany

Received May 21, 2013; Revised August 16, 2013; Accepted August 27, 2013

ABSTRACT

***De novo* motif discovery has been an important challenge of bioinformatics for the past two decades. Since the emergence of high-throughput techniques like ChIP-seq, ChIP-exo and protein-binding microarrays (PBMs), the focus of *de novo* motif discovery has shifted to runtime and accuracy on large data sets. For this purpose, specialized algorithms have been designed for discovering motifs in ChIP-seq or PBM data. However, none of the existing approaches work perfectly for all three high-throughput techniques. In this article, we propose Dimont, a general approach for fast and accurate *de novo* motif discovery from high-throughput data. We demonstrate that Dimont yields a higher number of correct motifs from ChIP-seq data than any of the specialized approaches and achieves a higher accuracy for predicting PBM intensities from probe sequence than any of the approaches specifically designed for that purpose. Dimont also reports the expected motifs for several ChIP-exo data sets. Investigating differences between *in vitro* and *in vivo* binding, we find that for most transcription factors, the motifs discovered by Dimont are in good accordance between techniques, but we also find notable exceptions. We also observe that modeling intra-motif dependencies may increase accuracy, which indicates that more complex motif models are a worthwhile field of research.**

INTRODUCTION

New high-throughput techniques such as ChIP-seq (1), ChIP-exo (2) and protein-binding microarrays (PBMs) (3) have dramatically increased the amount and quality of data that can be used for *de novo* motif discovery. ChIP-seq experiments determine binding regions of DNA-binding proteins *in vivo* by cross-linking protein and DNA, immunoprecipitating the targeted protein and sequencing the bound fragments. In case of ChIP-exo, the fragments are shortened by an exonuclease before sequencing. PBMs allow for measuring probe-specific binding affinity *in vitro* for a huge number of systematically chosen double-stranded probes. Despite the experimental differences, these approaches yield thousands of candidate binding regions together with a measure of confidence, which can be used for *de novo* motif discovery.

Ma *et al*. (4) provide an extensive comparison of *de novo* motif discovery tools capable of using ChIP-seq data, where ChIPMunk (5) and POSMO (4) are the best-performing tools closely followed by DME (6), DREME (7) and MEME (8). A detailed comparison of *de novo* motif discovery tools using PBM data is given by Weirauch *et al*. (9), where FeatureREDUCE emerges as top-performing algorithm. However, there is no tool that works well for data from both experimental techniques (9). For ChIP-exo data, no specialized tool is currently available, and research resorts to well-established algorithms from the pre-NGS era (2).

The lack of a universally applicable approach hampers the integration of data from different techniques and complicates the comparison of the resulting motifs, e.g. between *in vivo* and *in vitro* binding. Hence, we propose Dimont, a general approach for probabilistic discriminative *de novo* motif discovery that is capable of handling ChIP-seq, ChIP-exo and PBM data.

The runtime of most probabilistic *de novo* motif discovery tools is mainly determined by iteratively evaluating the likelihood. As the positions of the binding sites within the target sequences are unknown (hidden variables), these tools need to consider all admissible binding site positions for evaluating the likelihood, which has a decisive influence on runtime. One approach to circumvent this problem is to resort to *k*-mer enumeration methods like

*To whom correspondence should be addressed. Tel: +49 3946 47 510; Fax: +49 3946 47 500; Email: jens.keilwagen@jki.bund.de

POSMO (4), which yields a competitive runtime even on large data sets. Dimont implements an alternative approach that allows for adhering to probabilistic methods using the popular 'zero or one occurrence per sequence' (ZOOPS) model of many *de novo* motif discovery tools (8,10–13) while achieving acceptable runtimes. Dimont uses that only a few binding sites are buried within long target sequences. In most probabilistic approaches, this results in a big discrepancy between the number of finally predicted binding sites and the number of positions that need to be evaluated for computing the likelihood, and in wasting a considerable amount of runtime during training.

Hence, we only consider those positions contributing the most to the likelihood of a target sequence (Figure 1). During optimization, we dynamically determine the positions to be evaluated keeping the learning scheme flexible to adapt to the positions of potential binding sites. This acceleration scheme allows for using all ChIP binding regions or all PBM probe sequences for *de novo* motif discovery instead of limiting the input data to a fixed number of (high-confidence) sequences (14).

As peak occupancies or probe intensities contain valuable information for motif discovery, Dimont converts these to soft labels reflecting the *a priori* probability of a sequence being bound. These soft labels are used for learning parameters by a weighted variant (15) of the discriminative maximum supervised posterior principle (16,17).

In previous studies, the complexity of motif models was limited mostly owing to the limited amount of data. For this reason, simple models including consensus sequences as well as position weight matrices and sequence logos as their graphical representation are widespread. However, due to the enormous amount of high-throughput data, more complex models including inhomogeneous Markov models of higher order, which have been proven advantageous for other binding sites (18,19), can be used for *de novo* motif discovery and prediction of transcription factor binding sites. Hence, we include the capability of learning higher-order inhomogeneous Markov models into Dimont.

We implement Dimont within the open-source Java library Jstacs (20). We provide a Dimont web server at http://galaxy.informatik.uni-halle.de and a stand-alone command line application at http://www.jstacs.de/index.php/Dimont.

MATERIALS AND METHODS

The input data of Dimont are DNA sequences $\underline{x} = x_1, \ldots, x_L$ where each symbol x_ℓ is from the DNA alphabet $\Sigma = \{A,C,G,T\}$. Each of the sequences is assigned some measure of evidence that reflects how likely this sequence is bound by the transcription factor of interest. In case of ChIP-seq and ChIP-exo data, such a measure is the number of reads or fragments under a ChIP peak, often termed 'peak statistic' or 'peak occupancy'. For PBM data, such a measure is the signal intensity of the probe sequence on the microarray.

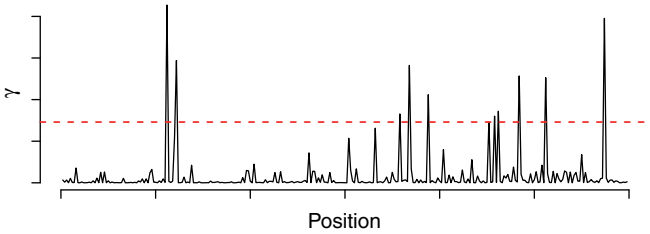

Figure 1. Normalized likelihood profile of a sequence. The red dashed line visualizes the threshold that is used to accelerate the algorithm. All positions with peaks above the threshold are included in $\mathcal{L}$, and all remaining positions are not used for evaluating the likelihood.

In the following, we assume that high-confidence sequences, i.e. those with a high peak statistic or a high signal intensity, contain a binding site of the motif of interest with substantially higher probability than low-confidence sequences. Hence, we transform these measures to probabilities that reflect how likely a sequence is bound by the transcription factor as explained in '*Soft labels from peak statistics and signal intensities*' section. For ChIP-seq and ChIP-exo data, we additionally assume that binding sites of the targeted transcription factor occur clustered around the centers of the ChIP peaks. Hence, we use a non-uniform position distribution over the binding site positions in the Dimont model, which we introduce in '*Dimont models and objective function*'. In subsequent sections, we describe how we accelerate the optimization of the parameters of the Dimont model, we outline the complete Dimont algorithm, and we introduce the performance measures and data sets used in the case studies of this article.

Soft labels from peak statistics and signal intensities

We map the peak statistics of ChIP data and the signal intensities of PBM data to soft labels that reflect the probability assumed *a priori* of being bound by the targeted factor. For this reason, we refer to the probabilty of being bound as 'foreground probability' and to the converse probability as 'background probability'.

Here, we propose a mapping that is based on the ranks of the signals within a data set. We denote as r_n the rank of the n-th sequence $\underline{x}_n$ in the data set. Let $m = \max_n\{r_n\}$ be the maximum rank, and let $h_n = \frac{r_n}{m}$ be the relative rank. We set q to the *a priori* fraction of sequences that receives a foreground probability greater than 0.5, and we refer to q as 'weighting factor'. The value of q can be adapted to the characteristics of the data, for instance, the significance level of accepted ChIP-seq peaks. In general, it is reasonable for any data source to also include low-confidence sequences into the input data to preserve the discriminative nature of Dimont. In our studies, we use $q = 0.2$ for ChIP data and $q = 0.01$ for PBM data. We define the foreground probability of sequence $\underline{x}_n$ as

$$w_n^{fg} := \frac{1}{1+\frac{h_n}{1-h_n}\cdot\frac{1-q}{q}}, \quad (1)$$

and the background probability as $w_n^{bg} := 1 - w_n^{fg}$. For simplicity reasons, we refer to the sequences in

conjunction with the foreground probability as 'foreground' and to the same sequences in conjunction with the background probability as 'background'.

Dimont models and objective function

Dimont is based on the popular ZOOPS model used in many *de novo* motif discovery tools (8,10–13). In Dimont, the motif model is a uniform mixture model over the DNA strands using an inhomogeneous Markov model of user-specified order, which includes the position weight matrix (PWM) model (21,22) for order 0 and the weight array matrix model (18,19) for order 1. We give a detailed definition of the likelihood of the motif model in Section 1 of the Supplementary Material.

In addition, we use a non-uniform position distribution $P(\ell)$ over all possible binding site positions relative to an anchor position. More specifically, we use a Gaussian distribution with given initial standard deviation of 75 around the anchor position for ChIP-seq, ChIP-exo and PBM data (details in Section 2 of the Supplementary Material).

For positions not covered by a binding site, we use a uniform distribution. Hence, the likelihood $P_{fg}(\underline{x}|\underline{\lambda})$ of an input sequence $\underline{x}$, given the ZOOPS model with parameters $\underline{\lambda}$ is defined as

$$P_{fg}(\underline{x}|\underline{\lambda}) = \frac{P(\text{motif} \mid \underline{\lambda})}{|\Sigma|^{L-w}} \sum_{\ell \in \mathcal{L}} P(\ell) P_M(x_\ell, \ldots, x_{\ell+w-1} \mid \underline{\lambda}) + \frac{1 - P(\text{motif} \mid \underline{\lambda})}{|\Sigma|^{L}} \quad (2)$$

where $P(\text{motif} \mid \underline{\lambda})$ denotes the *a priori* probability of observing a motif in a sequence, $\mathcal{L}$ is the set of admissible positions, initially set to $[1, L-w+1]$, and $P_M(x_\ell, \ldots, x_{\ell+w-1} \mid \underline{\lambda})$ denotes the likelihood of the motif model of width w. During optimization, we adapt $\mathcal{L}$ according to the acceleration scheme described in *Accelerated discriminative learning* section.

As background model $P_{bg}(\underline{x} \mid \underline{\lambda})$, we use either a uniform distribution or a homogeneous Markov model of order d.

To optimize the parameters of these models, we introduce a weighted variant (15) of the discriminative maximum supervised posterior principle (17,16,23) to *de novo* motif discovery, i.e.

$$\hat{\underline{\lambda}} = \underset{\underline{\lambda}}{\operatorname{argmax}} \sum_{n=1}^{N} \sum_{c \in \mathcal{C}} w_n^c \log \left(\frac{P(c \mid \underline{\lambda}) P_c(\underline{x}_n \mid \underline{\lambda})}{\sum_{\tilde{c} \in \mathcal{C}} P(\tilde{c} \mid \underline{\lambda}) P_{\tilde{c}}(\underline{x}_n \mid \underline{\lambda})} \right) + Q(\underline{\lambda} \mid \underline{\alpha}), \quad (3)$$

where $\mathcal{C} = \{fg, bg\}$ is the set of classes, and $Q(\underline{\lambda} \mid \underline{\alpha})$ denotes the prior on the parameters $\underline{\lambda}$ given hyper-parameters $\underline{\alpha}$. In case of Dimont, this prior is a transformed product-Dirichlet prior (23) using BDeu hyper-parameters (24,25) based on an equivalent sample size of 4 for the foreground class and $4 \cdot \frac{1-q}{q}$ for the background class. Parameter optimization is performed numerically using conjugate gradients.

Accelerated discriminative learning

We achieve an acceleration of parameter optimization by two general ideas. First, we perform a pre-optimization of parameters using a 'reduced data set' containing the highest-confidence sequences of foreground and background class (Hence, these sequences also correspond to the lowest-confidence sequences of the alternative class). To this end, we select the 30% of the sequences, but not >1000 sequences in total, obtaining the most extreme probabilities w_n^{fg} and w_n^{bg}, respectively. We select these sequences such that the proportion of foreground and background probabilities is approximately identical to the full data set by successively adding sequences with the highest w_n^{fg} and w_n^{bg}, respectively.

Second, we observe that only few binding sites are detected within long target sequences as exemplarily depicted in Figure 1. A large proportion of runtime is wasted while evaluating the likelihood of the motif model for positions that will never be predicted as potential binding sites. Hence, we only use the most relevant positions corresponding to the largest summands in Equation (2) instead of computing all terms. For this reason, we compute and normalize all summands of Equation (2) for each sequence $\underline{x}$ yielding

$$\gamma_\ell := \frac{P(\ell) P_M(x_\ell, \ldots, x_{\ell+w-1} \mid \underline{\lambda})}{\sum_{\tilde{\ell}=1}^{L-w+1} P(\tilde{\ell}) P_M(x_{\tilde{\ell}}, \ldots, x_{\tilde{\ell}+w-1} \mid \underline{\lambda})}. \quad (4)$$

We then rank the positions ℓ by γ_ℓ in descending order. This rank is different from the rank r_n according to the peak statistics or signal intensities, respectively, as given by the biological experiment. Here, the rank reflects the prediction due to the statistical model. Subsequently, we select in descending order a set of relevant positions $\mathcal{L}$ until $\sum_{\ell \in \mathcal{L}} \gamma_\ell \geq 0.5$, and we refer to this threshold as 'likelihood cutoff'. During numerical optimization, we determine $\mathcal{L}$ at the beginning of each iteration using the current set of parameters $\underline{\lambda}$. Evaluating the likelihood of Equation (2) in the numerical optimization, we only use the positions in $\mathcal{L}$.

The Dimont algorithm

In the following, we describe the Dimont algorithm step by step.

Pre-processing

We read the input sequences including peak statistics or probe intensities, which we convert to soft labels.

Initialization

For initializing the motif model, we first enumerate all 7mers that occur in the reduced data set. We then rank these 7mers by $\log(n_{fg}) \cdot n_{fg}/n_{bg}$, where n_{fg} is the sum of the foreground probabilities w_n^{fg} of all sequences $\underline{x}_n$ containing the current 7mer at least once, and n_{bg} is the corresponding sum of the background probabilities w_n^{bg}. We filter the ranked 7mers by excluding redundant variants, which have a Hamming distance of <2 to better-ranked 7mers.

Of the ranked and filtered 7mers, we select the top 50 7mers and use each of these to initialize the core of a motif

model of initial width w such that the central positions obtain a probability of 0.9 for the corresponding nucleotide in the 7mer and a probability of $\frac{0.1}{3}$ for the remaining nucleotides. The bordering positions are assigned a uniform distribution. We then evaluate the conditional likelihood, i.e. Equation (3) without the prior term, and choose the top m initial motifs with respect to conditional likelihood.

Pre-optimization

For each of the m initial motifs, we optimize the parameters according to Equation (3) on the reduced data set using the accelerated optimization described in the previous section. We then rank the resulting motifs by the supervised posterior achieved in the optimization.

Filtering motifs

Initialization and pre-optimization may result in redundant motifs, e.g. shifted variants or reverse complementary motifs. To reduce runtime, we filter such redundant motifs before the final optimization. We consider two motifs redundant if their score profiles, i.e. their γ_ℓ for all positions $\ell = 1, \ldots, L - w+1$ show a Pearson correlation greater than 0.3, averaged over all sequences in the reduced data set. During this filtering step, we allow for shifts of the score profiles up to w in both directions. If two motifs are considered redundant, we keep the motif variant achieving the larger supervised posterior.

Final optimization

For each of those motifs that remain after the filtering step, we optimize the parameters with respect to Equation (3) on the complete input data set. Subsequently, we compute the Kullback–Leibler divergence (26) between the marginal distribution at each motif position and the nucleotide composition of the complete data set. We remove bordering motif positions as long as Kullback–Leibler divergence is below 0.2. If Kullback–Leibler divergence exceeds 0.8 for a bordering position, we expand the motif by one additional position. We then adjust the standard deviation of the position distribution, and finally optimize the parameters with respect to Equation (3) on the complete input data set. Again, we rank the resulting motifs by the supervised posterior achieved in the optimization on the complete data set.

Post filtering

We finish the Dimont algorithm with a final filtering step in analogy to pre-optimization to eliminate redundantly reported motifs.

Default parameters

As default parameters of Dimont, we suggest (i) a motif model of order of 0, i.e. a PWM model; (ii) a uniform background model; (iii) a weighting factor of $q = 0.2$; (iv) an initial motif width of $w = 15$; and (v) $m = 20$ pre-optmization runs. We use these default parameters throughout this article if not stated otherwise.

Performance measures

For evaluating the performance of *de novo* motif discovery predictions, several measures have been used. For PBM data, we stick to the area under the receiver-operating characteristic curve (AUC-ROC) and Pearson correlation, as these have been used as final performance measures in the DREAM5 challenge (9). Pearson correlation is sensitive to monotone transformations of the predicted scores, while AUC-ROC is insensitive to such transformations. For maximizing the Pearson correlation, we search an adequate transformation,

$$f(r \mid c) = \frac{c \cdot r}{1 + \mid c \cdot r \mid}, \tag{5}$$

where r is the predicted score, namely, the likelihood ratio, and c is a free parameter. We optimize c to maximze the Pearson correlation on the training data and use this optimal value to transform the likelihood ratios of the test data. For computing AUC-ROC, probe sequences with a mean signal intensity >4 standard deviations above the experiment average are assigned to the positive class, and all other probe sequences are assigned to the negative class (9).

Comparing the results of Dimont to other tools and between experiments, we use sequence logos as proposed by Ma *et al.* (4), the normalized Euclidean distance as proposed by Linhart *et al.* (27) and AUC-ROC.

Data

ChIP-seq data

We obtain the ChIP-seq peaks (centers and peak statistics) of the 26 ChIP-seq data sets compiled by Ma *et al.* (4) from original publications (28–34). For the comparison of ChIP-seq and PBM data, we additionally obtain the ChIP-seq peaks of Foxo1 (GSM546525, (35)), GATA4 (GSM558904, (36)), Tcf3 (GSM915177, unpublished) Tbx5 (GSM558908, (36)) and Tbx20 (GSM734426, (36)) from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), and of Nr5a2 (SRP001796, (37)) from the hmChIP database ((38), http://jilab.biostat.jhsph.edu/database/cgi-bin/hmChIP.pl).

For each of these data sets, we download the genome sequences of the corresponding species and genome version (hg18, mm8, mm9, dm3) from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/downloads.html). For each ChIP-seq peak, we extract 1000 bp of genomic sequence centered around the given peak summit and annotate these sequences with the corresponding peak statistic.

ChIP-exo data

We obtain the ChIP-exo peaks (peak coordinate and occupancy) from the supplement of Rhee and Pugh (2). For CTCF, we download the human genome sequences (hg18) from the UCSC Genome Browser. In case of the three yeast data sets, we obtain the yeast genome (build 19-Jan-2007) from the Saccharomyces Genome Database (http://www.yeastgenome.org/download-data). For each ChIP-exo peak, we extract, based on CW distance, 200 bp (CTCF) or 100 bp (yeast factors) of genomic sequence centered around the given peak center, and we annotate these sequences with the corresponding peak occupancy.

PBM data

We obtain the 40 tuning, the 66 training and the 66 test PBM data sets of DREAM5 challenge2 (http://wiki.c2b2.columbia.edu/dream/index.php/D5c2). For the comparison of ChIP-exo and PBM data, we additionally obtain PBM data sets for Phd1 (UP00351) and Rap1 (UP00321) from the UniPROBE database (39,40) (http://thebrain.bwh.harvard.edu/uniprobe/). Of each probe sequence, we extract the first 40 bp, comprising 35 unique base pairs and 5 bp of linker sequence.

In case of the PBM data sets of DREAM5, we follow the proposal of Weirauch *et al.* (9) and use the mean signal intensities after spatial detrending and quantile normalization.

RESULTS

Runtime

We assess the runtime of Dimont on all data sets considered in this article on a standard laptop (Intel Core i7, ULV, dual core, 2 Ghz) using standard parameters. In Figure 2, we plot the runtime of Dimont against the size of the input data set for different types of input data. For ChIP-seq data sets comprising sequences of length 1000 bp, we observe runtimes of ~5 min for medium sized data sets. On the largest ChIP-seq data set containing 73 795 sequences of length 1000 bp, Dimont runs for 1 h 15 min. Without the speed-up strategy described in *Accelerated discriminative learning* section, runtime would increase by a factor of 5 to 29 as shown for selected data sets, namely, KNI (504 sequences), c-Myc (3 413 sequences), KR2 (5 793 sequences) and FoxA2 (11 461 sequences). We give a detailed overview of runtime dependency on the speed-up strategy and motif order in Supplementary Figures S1–S3.

For ChIP-exo data sets comprising sequences of length 200 in case of CTCF and sequences of length 100 in case of the yeast data sets, runtime decreases substantially, and Dimont reports a motif after at most 5 min.

In case of the PBM data containing ~40 000 probe sequences of length 40 bp per data set, Dimont runs for 2–8 min.

ChIP-seq

In a first case study, we assess Dimont using default parameters on the 26 ChIP-seq data sets of Ma *et al.* (4). In Figure 3, we present exemplary motifs for three of the factors considered, while the motifs reported for all data sets are available in Supplementary Figure S4. In addition to a visual comparison of the motifs discovered to those from the literature, we consider the normalized Euclidean distance (27) between the two motifs as a measure for their similarity.

The motif of FoxA2 discovered by Dimont closely resembles the motif reported in the Jaspar database (41) with clear consensus GTAAACA (normalized Euclidean distance $d = 0.06$). The motif of Tcfcp2l1 is also recovered well by Dimont ($d = 0.12$), although minor differences are visible: the strength of conservation at some positions differs between the motif reported by Dimont and that

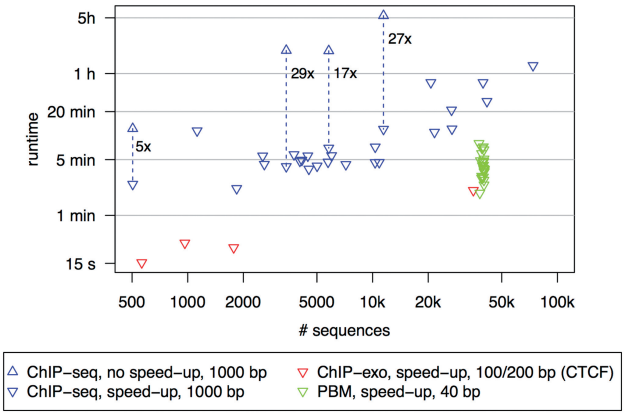

Figure 2. Runtime evaluation of Dimont on the data sets used in this article. We consider all ChIP-seq data sets (blue), ChIP-exo (red) and PBM (green) data sets used in this article. Upright triangles represent the runtime without the speed-up strategy, whereas reversed triangles represent the runtime using the speed-up strategy. Runtime decreases by a factor of 5 to 29 due to the speed-up strategy.

of Jaspar. In addition, Dimont includes two additional positions with a slight preference for A into the motif, while the last conserved G, present in the Jaspar motif, is omitted. The latter might be an effect of the strand model of Dimont combined with the roughly palindromic structure of Tcfcp2l1.

The motif of KNI ($d = 0.20$) is one of three motifs that are discovered from ChIP-seq data exclusively by Dimont (c.f. Supplementary Figure S4, Table 1). We find that the consensus of the Jaspar motif (AAANTAGAGCA) fits the motif discovered by Dimont. However, we find two notable differences between the two motifs. First, the sequence of As at the 5′ end of the motif is more conserved in the Jaspar motif. Second, we find mildly conserved Gs at positions 4 and 12 of the motif reported by Dimont, which are not present in the Jaspar motif.

We assess the performance of Dimont on all data sets of Ma *et al.* (4) by counting the number data sets for which Dimont successfully discovers the known motif for the targeted transcription factor. We define a discovery successful *iff* the normalized Euclidean distance between the predicted motif and the motif described in the literature (4,33,34,41,42) is smaller than 0.25. We give an overview of this assessment in Table 1, and we additionally include the number of motifs correctly discovered by POSMO (4), MEME (8), DME (6), ChIPMunk (5), HMS (42) and DREME (7) as reported by Ma *et al.* (4). All motifs discovered by Dimont are presented in Supplementary Figure S4.

We find by comparing the discovered motifs to the literature using the normalized Euclidean distance that Dimont discovers all 26 motifs. As reported by Ma *et al.* (4), POSMO and ChIPMunk discover 23 motifs; MEME, DME and DREME discover 22 motifs; and HMS discovers 12 motifs. Three motifs (CAD, E2f1 and KNI) are discovered only by Dimont but by none of the previous approaches.

Considering the average rank of correct predictions, we find that for 20 of the 26 data sets, Dimont reports the correct motif on rank 1. For the remaining six data sets

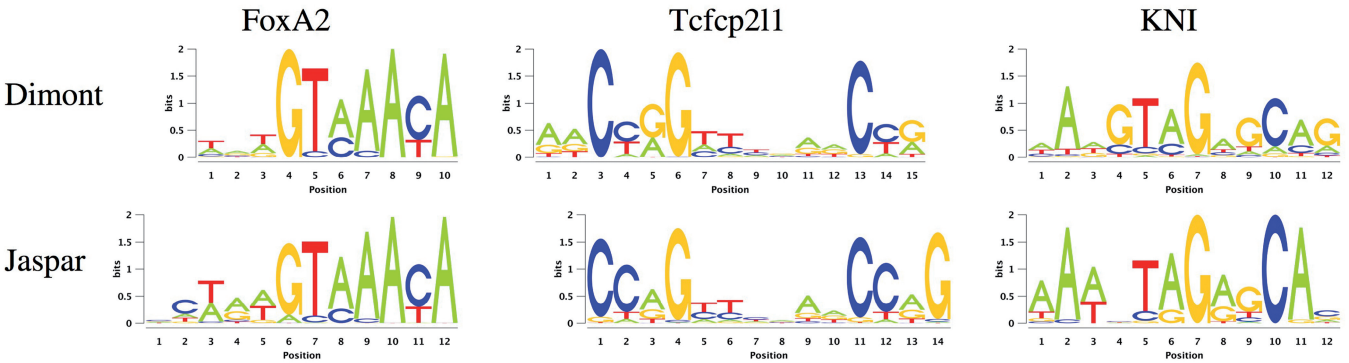

Figure 3. Three exemplary motifs discovered by Dimont on the FoxA2, Tcfcp2l1 and KNI data sets of Ma *et al.* (4) compared with the corresponding motifs from the Jaspar database.

Table 1. Number of motifs successfully discovered by Dimont on the data sets compiled by Ma *et al.* (4) compared with the results of POSMO, ChIPMunk, MEME, DME, DREME and HMS

Algorithm	Total successes	Average rank
Dimont	26	1.23
POSMO	23	1.00
ChIPMunk	23	1.00
MEME	22	1.32
DME	22	1.45
DREME	22	1.45
HMS	12	1.00

We define a discovery successful *iff* the normalized Euclidean distance between the predicted motif and the motif described in the literature is smaller than 0.25.

(CAD, GT, KNI, KR1, KR2 and Nanog), Dimont reports the correct motif only on rank 2. Scrutinizing such cases (Supplementary Figure S4), we find that the first motif reported for Nanog does not show a clear similarity to other known motifs. For KNI, Dimont reports the binding motif of CAD on rank 1, which can be explained by substantial co-binding of KNI and CAD (32). For four *Drosophila melanogaster* data sets (CAD, GT, KR1, KR2), the first motif reported by Dimont is almost identical having consensus CAGGTAG. The same motif is also discovered by Dimont as a second motif for the HB1 and BCD data sets. This motif is bound by the Zelda (ZLD) transcription factor, a member of the so-called TAGteam (43). ZLD has been reported to play a key role in transcriptional activation during maternal-to-zygotic transition, and regions bound by ZLD in early development are later occupied by several specific transcription factor including BCD, CAD, GT, KR and HB (44).

In summary, Dimont discovers all motifs of the ChIP-seq data sets compiled by Ma *et al.* (4), including three motifs that are not found by previous approaches. For the majority of data sets, Dimont returns the correct motif at rank 1, whereas rank 2 for the remaining data sets can often be explained by biological phenomena.

ChIP-exo

In a second case study, we investigate the capability of Dimont to discover motifs in ChIP-exo data. To this end, we consider four of the five ChIP-exo data sets compiled by Rhee and Pugh (2), human CTCF and Rap1, Reb1 and Phd1 from *Saccharomyces cerevisiae*. We exclude the Gal4 data set, as it contains only 15 binding regions.

We present the motifs reported by Dimont using default parameters for the yeast data sets in Figure 4. The motif discovered for Rap1 closely resembles the core of the 'telomeric' motif of Rap1 found by Rhee and Pugh (2) and is an extended variant of the motif reported in Jaspar. In case of Reb1, the consensus TACCCG of the discovered motif is identical to the previously reported Reb1 consensus (2,45) and highly similar to the Jaspar motif. For Phd1, Dimont finds a motif highly similar to the Phd1 motif discovered by Zhu *et al.* (39) from PBM data and to the Phd1 motif reported in Jaspar. Notably, this motif has not been discovered from these ChIP-exo data by Rhee and Pugh (2) using MEME for *de novo* motif discovery.

For the human insulator CTCF, ChIP-exo as well as ChIP-seq data are available. We show a comparison of the motifs discovered by Dimont from the ChIP-seq and ChIP-exo data sets to the motif present in Jaspar in Figure 5. All three motifs are highly similar, whereas the level of conservation slightly differs for some positions.

In summary, Dimont discovers the binding motifs of all four transcription factors from the ChIP-exo data sets considered.

Protein binding microarrays

In a third case study, we consider the applicability of Dimont to PBM data. To this end, we assess the performance of Dimont on the data provided by DREAM5 challenge2 (cf. *PBM data* section). In this challenge, the signal intensities of one PBM layout should be predicted based on the probe sequences and the signal intensities of all probes of another PBM layout. During the challenge, tuning data for both PBM layouts were provided for calibrating external parameters of the participating approaches. We use these tuning data to determine (i) the optimal order d of the background model and (ii) the optimal weighting factor q for PBM data, whereas the initial motif width and the number of pre-optimization runs are left at their default values (cf. *The Dimont*

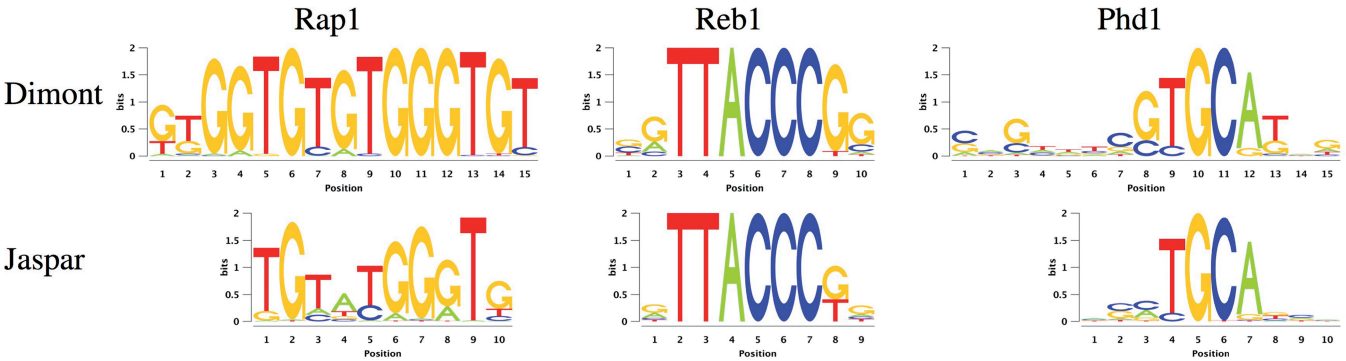

Figure 4. Motifs discovered by Dimont on three of the yeast ChIP-exo data sets of Rhee and Pugh (2) compared with the corresponding motifs from the Jaspar database.

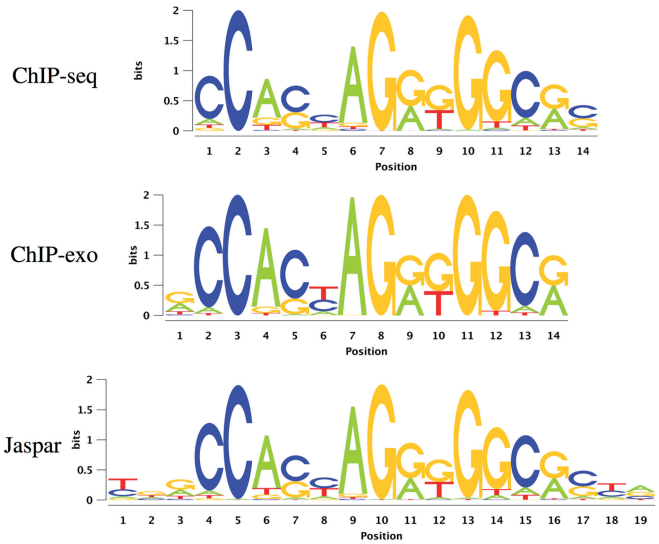

Figure 5. Motifs discovered by Dimont on the ChIP-seq and ChIP-exo data sets of the human insulator CTCF compared with the CTCF motif from the Jaspar database.

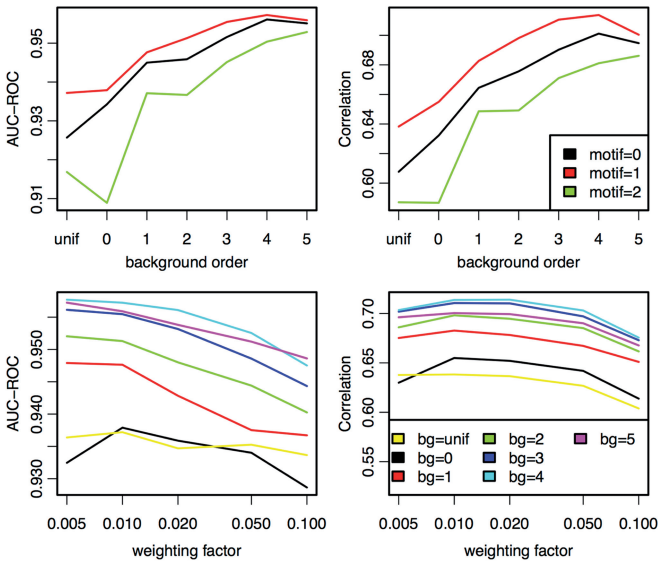

Figure 6. Influence of the choice of background order for different motif orders and weighting factor on the performance on the tuning data sets of the DREAM5 challenge. In the first row, we plot performance against background order for motif orders 0, 1 and 2 and a fixed weighting factor of 0.01. In the second row, we plot performance against weighting factors for a uniform background model and background orders 0 to 5, given a fixed motif order of 1.

algorithm section). We present the results of these analyses in Figure 6. Regarding the order of the background model, we find consistently for motif orders 0 to 2 that prediction performance as measured by AUC-ROC and Pearson correlation increases up to a background order of 4. From the first row of Figure 6, we also observe that motif order 1 (weight array matrix (WAM) model) performs consistently better than orders 0 and 2.

Hence, we fix the motif order to 1 in the second row of Figure 6 and investigate the influence of the weighting factor q on the predictions performance for different background orders. We find that for higher background orders, AUC-ROC increases with decreasing weighting factor. Considering Pearson correlation, a weighting factor of 0.01 performs slightly better than 0.005, whereas 0.02 reaches a comparable correlation for most background orders.

Allowing for model selection with regard to motif order, we choose for each data set the motif order yielding the maximum AUC-ROC on the training data set and test the prediction performance on the corresponding test data set. Doing so for the tuning data sets, the performance slightly increases yielding an AUC-ROC of 0.958 and a Pearson correlation of 0.714.

Given these results on the tuning data, we fix the background order to 4 and the weighting factor to 0.01 in the following analyses on the DREAM5 training and test data. We train Dimont for motif orders 0 to 2 on each of the 66 training data sets and allow for selection of motif order on the training data. Following the proposal of Weirauch *et al.* (9), we consider the average Pearson correlation *cc* and the average AUC-ROC *roc* over all 66 test data sets, and we compute a final score as $(cc/0.696 + (roc - 0.5)/(0.949 - 0.5))/2$. Thereby, 0.696 is the maximum Pearson correlation, and 0.949 is the maximum AUC-ROC gained by any of the approaches considered by Weirauch *et al.* (9).

Table 2. Performance of Dimont on the DREAM5 data compared with the best approaches according to Weirauch *et al.* (9) for predicting PBM signal intensities from probe sequence as measured by Pearson correlation, AUC-ROC and a combined final score

Algorithm	Pearson corr.	AUC-ROC	Final
Dimont	0.695	**0.951**	**1.002**
FeatureREDUCE	0.693	0.949	0.997
Team_D	0.691	0.938	0.984
Team_E	**0.696**	0.906	0.952

In Table 2, we compare the prediction accuracy achieved by Dimont to that of the top performers according to Weirauch *et al.* (9), namely, FeatureREDUCE, Team_D and Team_E. The maximum Pearson correlation of 0.696 is gained by Team_E, whereas among the existing approaches, the maximum AUC-ROC of 0.949 is gained by FeatureREDUCE. We find that Dimont achieves a Pearson correlation of 0.695, which is slightly greater than the Pearson correlation of the top performer FeatureREDUCE but slightly smaller than the Pearson correlation gained by Team_E. Considering AUC-ROC, Dimont yields a slightly greater AUC-ROC than all of the existing approaches considered. However, because of the large variation between the different data sets, neither of these improvements can be considered significant.

Combining Pearson correlation and AUC-ROC, Dimont yields a greater final score than FeatureREDUCE, Team_D, Team_E and all other approaches considered by Weirauch *et al.* (9).

As model selection with regard to motif order further increases the prediction performance of Dimont, we consider the selected model orders for different families of transcription factors, and we give a complete list of chosen model orders in Supplementary Table S1. For most families, we do not find a clear preference for a specific motif order. Notable exceptions are the AT_hook family, which appears to profit from second-order dependencies, the bHLH and nuclear receptor families showing a preference for motif order 1, and the C2H2 zinc finger family, which shows a slight shift to motif order 0 compared with all transcription factors. Motif order 0 is chosen for less than one-third of the data sets, whereas higher motif orders are preferred for more than two-thirds of the data sets.

Comparison of de novo motif discovery using different experimental techniques

Owing to general applicability of Dimont to ChIP-seq, ChIP-exo and PBM data demonstrated in the previous sections, we have the opportunity to investigate the consistency of the discovered motifs between *in vitro* and *in vivo* binding and between different technologies. To this end, we consider all transcription factors for which on the one hand PBM data and on the other hand ChIP-seq or ChIP-exo data are available, and CTCF for a ChIP-seq/ChIP-exo comparison.

In a first study, we run Dimont on the PBM data set and the corresponding ChIP data set using a PWM model and the standard parameters for each technology (ChIP-seq/ChIP-exo: uniform background, $q = 0.2$, $w = 15$, $m = 20$; PBM: background order $d = 4$, $q = 0.01$, $w = 15$, $m = 20$) and compare the resulting binding motifs. We present the results of this study in Figure 7. For many data sets, namely, Esrrb, Foxo1, Gata4 and Zfx, we obtain largely similar motifs for both, PBM and ChIP-seq/ChIP-exo data. This indicates than *in vitro* binding assays like PBMs are a valuable technique to determine binding specificities that are also valid *in vivo*. For Nr5a2, Phd1, Rap1 and Tcf3, we find minor differences between the PBM and the ChIP-seq/exo motif, which are basically different levels of conservation and differences in the number of flanking positions. We observe the greatest differences for the two T-box motifs, namely, Tbx5 and Tbx20. The PBM motifs of Tbx5 and Tbx20 are similar, both having consensus TNACACCT, and agree with *in vitro* T-box motifs from the literature (46,47). The ChIP-seq motifs for both factors differ substantially from their PBM counterparts and from each other. Although the reason for this observation remains unclear, a similar *in vivo* motif of Tbx20 and a similar discrepancy between *in vitro* and *in vivo* binding of Tbx20 has been reported before (47), which might indicate similar effects for other T-box factors including Tbx5. An alternative explanation might be that *in vivo* Tbx5 co-binds with another factor enriched in the top ChIP-seq peaks. However, increasing q up to 0.6 does not result in a different motif, although a greater number of sequences are considered to be bound.

In a second study, we consider classification across technologies as an additional indication of the compliance of *in vitro* and *in vivo* binding. In case of PBM data, we use the partitioning into positive and negative probe sequences proposed in the DREAM5 challenge (9). For ChIP-seq and ChIP-exo data, the positive class contains the sequences around the top 500 ChIP peaks, and the negative class comprises 10 shuffled variants of each positive sequence preserving di-nucleotide content. We assess the classification performance across technologies and for model order 0 and 1 in a 10-fold cross-validation (details given in Section 6 of the Supplementary Material). For the assessment in each iteration of a cross-validation run, we use only the motif reported by Dimont at rank 1.

We use the Dimont classifiers obtained on the ChIP and PBM training data to classify both the PBM and ChIP data sets for the same transcription factor. For PBM data, we train the classifier using background order 4 as before but replace the background model by a uniform distribution for testing to eliminate influences aside the motif model on classification performance. We present the results of this cross-validation in Table 3.

For Esrrb, Foxo1, Gata4, Nr5a2 and CTCF, the classifiers applied to data from a different technology than used for training achieve a performance that is comparable with the intra-technology case. In case of Tbx20 and Tbx5, we observe a considerably decreased performance in at least one direction of the cross-technology comparison, a result that is consistent with the previous statements on the motif level. Although the PBM classifiers for Tcf3 and Zfx show a decreased AUC-ROC for ChIP-seq test data,

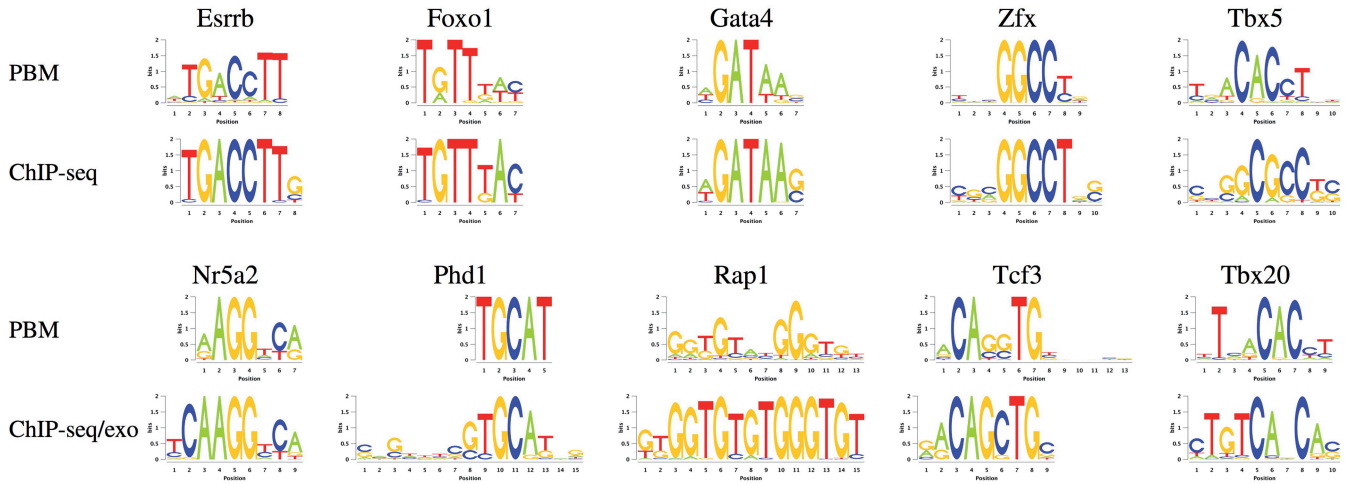

Figure 7. Comparison of the motifs discovered by Dimont using PBM and ChIP-seq or ChIP-exo data. For Esrrb, Foxo1, Gata4 and Zfx, we obtain largely similar motifs for PBM and ChIP-seq/ChIP-exo data, whereas we find minor differences for Nr5a2, Phd1, Rap1 and Tcf3. In case of Tbx5 and Tbx20, the motifs discovered from PWM and ChIP-seq data differ substantially.

Table 3. Mean AUC-ROC of a 10-fold cross-validation

Factor	Order 0	Order 1	Order 0	Order 1	Order 0	Order 1	Order 0	Order 1
	Test(ChIP-seq)				Test(PBM)			
	Train(ChIP-seq)		Train(PBM)		Train(ChIP-seq)		Train(PBM)	
Esrrb	0.922	**0.930***	**0.901***	0.885	0.896	**0.908**	0.861	**0.906***
Foxo1	0.746	**0.768***	0.752	**0.794***	**0.902***	0.868	0.957	**0.962**
Gata4	0.787	**0.807***	0.739	**0.777***	0.974	0.974	**0.983***	0.979
Nr5a2	0.853	**0.858**	0.858	**0.866***	**0.910***	0.864	0.963	**0.965**
Tbx20	**0.772**	0.770	0.512	**0.524**	0.570	**0.691***	**0.994**	0.990
Tbx5	0.629	**0.634**	**0.604***	0.591	**0.808***	0.550	0.992	**0.993**
Tcf3	**0.929**	0.925	0.784	**0.807***	**0.973***	0.886	0.973	**0.977***
Zfx	**0.723**	0.719	0.556	**0.563**	**0.950***	0.942	**0.970**	0.967
	Test(ChIP-seq)				Test(ChIP-exo)			
	Train(ChIP-seq)		Train(ChIP-exo)		Train(ChIP-seq)		Train(ChIP-exo)	
CTCF	**0.882**	0.881	0.800	**0.806***	**0.909**	0.907	0.877	**0.879**
	Test(ChIP-exo)				Test(PBM)			
	Train(ChIP-exo)		Train(PBM)		Train(ChIP-exo)		Train(PBM)	
Phd1	**0.634**	0.621	0.632	**0.661***	0.786	**0.889***	**0.962***	0.957
Rap1	**0.781***	0.766	0.800	**0.819***	**0.758***	0.727	0.823	**0.837**

We train Dimont on ChIP-seq, PBM or ChIP-exo data and apply each of the resulting classifiers to each of the available data sets for the same transcription factor. Comparing AUC-ROC for motif orders 0 and 1, the maximum is displayed in bold face, and significant differences are marked with an asterisk.

the ChIP-seq classifiers for these data sets yield a comparable performance on the PBM test data as for the ChIP-seq test data. In both cases, one explanation might be the low number of conserved motif positions (cf. Figure 7), which leads to a large number of random hits in the shuffled negative sequences. For Phd1 and Rap1, the ChIP-exo classifiers yield lower AUC-ROC values on the PBM data than the PBM classifiers, whereas the converse combinations yield a classification that is comparable with the ChIP-exo classifiers.

In the previous section, we observed that increasing the motif order increases the prediction performance of Dimont for PBM data. The existence of PBM, ChIP-seq and/or ChIP-exo data for the same transcription factors allows for investigating whether this observation is due to artifacts of PBM data or due to true dependencies between adjacent positions of transcription factor binding sites. In Table 3, we find that classifiers trained on PBM data and applied to ChIP data often achieve a greater classification performance for motif order 1 than for motif order 0, whereas the opposite tendency can be observed for the classifier trained on ChIP data and applied to PBM data. One explanation might be that the systematic design of PBMs combined with the large number of probe sequences allows for capturing true dependencies between adjacent positions, whereas the dependencies learned from ChIP-seq data are also influenced by general dependencies in the long input sequences. An alternative explanation could be that different modes of binding exist for several transcription factors, where only one of these modes is relevant for *in vivo* binding, but both are represented in PBM data. Such heterogeneities could be represented by higher order motif models, but not by PWMs. We study the dependencies discovered by Dimont for all

data sets, which show a significantly greater AUC-ROC for motif order 1 than for motif order 0 for at least one combination of training and test data sets in Supplementary Figure S5–S13, and we compare the dependencies detected by Dimont to those detected by diChIPMunk (48) in Section 6 of the Supplementary Material.

DISCUSSION

New high-throughput techniques including ChIP-seq, ChIP-exo and PBMs have greatly increased the quality and amount of data that are available for *de novo* motif discovery. Specialized tools have been developed for discovering motifs in ChIP-seq data, and other tools have been developed for discovering motifs in PBM data. However, none of the current tools work perfectly across all of these techniques, which hampers integration of data from different techniques and cross-technology comparison of the resulting motifs.

Hence, we developed Dimont, a tool for *de novo* motif discovery from ChIP-seq, ChIP-exo and PBM data using an accelerated discriminative learning scheme. We test Dimont on a collection of 26 ChIP-seq data sets and observe that Dimont discovers all of the expected motifs, where three of these motifs could not be discovered by any previous approach. Hence, we may state that Dimont is currently one of the best-performing approaches for *de novo* motif discovery from ChIP-seq data. Applying Dimont to ChIP-exo data sets of three yeast factors and human CTCF, the discovered motifs are in well accordance to the literature. We also assess the performance of Dimont on the PBM data of DREAM5 challenge 2 and find that Dimont predicts signal intensities from PBM probe sequence with greater accuracy than previous approaches. Hence, we may state that Dimont is currently one of the best-performing approaches for predicting PBM intensity values from probe sequence. Against the background of these three benchmark studies, we may state that Dimont is a general approach for fast and accurate *de novo* motif discovery from ChIP-seq, ChIP-exo and PBM data. Although the runtime required by Dimont is greater than the runtime of the currently fastest approach, POSMO (4), we consider a maximum runtime of 1 h 15 min and a typical runtime of <10 min acceptable after days or weeks of wet-laboratory work.

We further investigate whether motifs discovered by Dimont from *in vitro* and *in vivo* data can be transferred from one technique to the other by comparing the discovered motifs and by cross-technology classification. For most transcription factors, we find a good generalization of the motifs discovered by Dimont, which indicates that *in vitro* experiments often yield motifs that are also valid for *in vivo* binding. However, we also observe substantial differences between *in vitro* and *in vivo* binding for two transcription factors, namely, Tbx5 and Tbx20.

For PBM data, we also observe that using an inhomogeneous Markov model of order 1 instead of the popular PWM model substantially increases prediction performance. We investigate whether this finding can also be transferred to ChIP-seq or ChIP-exo data. Indeed, we observe that increasing the motif order to 1 for *de novo* motif discovery from PBM data increases classification accuracy on PBM as well as ChIP-seq and ChIP-exo data in the majority of cases.

These findings indicate that with the increased amount of data due to current high-throughput techniques, motif models capturing dependencies between motif positions may be of great value for predicting transcription factor binding sites, especially for predicting *in vivo* binding sites given *in vitro* training data.

As Dimont is implemented in the open-source Java library Jstacs (http://www.jstacs.de), new models capturing such dependencies can flexibly be implemented and easily integrated into Dimont by advanced users.

AVAILABILITY

For instant use, we also provide a Dimont web server at http://galaxy.informatik.uni-halle.de and a stand-alone command line application at http://www.jstacs.de/index.php/Dimont.

ACKNOWLEDGEMENTS

The authors thank Matti Annala, Joseph Corbo, Timothy R. Hughes, Ashwinikumar Kulkarni, Xiaotu Ma, Huck Hui Ng, Matthew T. Weirauch and Michael Q. Zhang for providing data and scripts, and for valuable discussions.

FUNDING

Ministry of Culture of Saxony-Anhalt [XP3624HP/0606T] and institutional budget funds. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Funding for open access charge: Institutional budget funds.

Conflict of interest statement. None declared.

REFERENCES

1. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, **316**, 1497–1502.
2. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408– 1419.
3. Berger,M.F. and Bulyk,M.L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.*, **4**, 393–411.
4. Ma,X., Kulkarni,A., Zhang,Z., Xuan,Z., Serfling,R. and Zhang,M.Q. (2012) A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Res.*, **40**, e50.
5. Kulakovskiy,I.V., Boeva,V.A., Favorov,A.V. and Makeev,V.J. (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–2623.
6. Smith,A.D., Sumazin,P. and Zhang,M.Q. (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc. Natl Acad. Sci. USA*, **102**, 1560–1565.
7. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.

8. Bailey,T.L. and Elkan,C. (1994) Fitting a Mixture model by expectation maximization to discover motifs in biopolymers. In: *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 28–36.
9. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
10. Ao,W., Gaudet,J., Kent,W.J., Muttumu,S. and Mango,S.E. (2004) Environmentally Induced Foregut Remodeling by PHA-4/FoxA and DAF-12/NHR. *Science*, **305**, 1743–1746.
11. Redhead,E. and Bailey,T. (2007) Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics*, **8**, 385.
12. Kim,N.K., Tharakaraman,K., Marino-Ramirez,L. and Spouge,J. (2008) Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. *BMC Bioinformatics*, **9**, 262.
13. Keilwagen,J., Grau,J., Paponov,I.A., Posch,S., Strickert,M. and Grosse,I. (2011) *De-Novo* discovery of differentially abundant transcription factor binding sites including their positional preference. *PLoS Comput. Biol.*, **7**, e1001070.
14. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.
15. Grau,J. (2010) Discriminative Bayesian principles for predicting sequence signals of gene regulation. *PhD Thesis*. Martin Luther University Halle–Wittenberg, Halle, Germany.
16. Cerquides,J. and de Mántaras,R.L. (2005) Robust bayesian linear classifier ensembles. In: *Proceedings of the 16th European conference on Machine Learning*. Springer-Verlag ECML'05, Berlin, Heidelberg, pp. 72–83.
17. Roos,T., Wettig,H., Grünwald,P., Myllymaki,P. and Tirri,H. (June, 2005) On Discriminative Bayesian Network Classifiers and Logistic Regression. *Mach. Learn.*, **59**, 267–296.
18. Zhang,M. and Marr,T. (1993) A weight array method for splicing signal analysis. *Comput. Appl. Biosci.*, **9**, 499–509.
19. Salzberg,S.L. (1997) A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput. Appl. Biosci.*, **13**, 365–376.
20. Grau,J., Keilwagen,J., Gohr,A., Haldemann,B., Posch,S. and Grosse,I. (2012) Jstacs: a Java Framework for statistical analysis and classification of biological sequences. *J. Mach. Learn. Res*, **13**, 1967–1971.
21. Stormo,G.D., Schneider,T.D., Gold,L.M. and Ehrenfeucht,A. (1982) Use of the'perceptron' algorithm to distinguish translational initiation sites. *Nucleic Acids Res.*, **10**, 2997–3010.
22. Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505–519.
23. Keilwagen,J., Grau,J., Posch,S. and Grosse,I. (2010) Apples and oranges: avoiding different priors in Bayesian DNA sequence analysis. *BMC Bioinformatics*, **11**, 149.
24. Buntine,W.L. (1991) Theory Refinement of Bayesian Networks. *Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, pp. 52–62.
25. Heckerman,D., Geiger,D. and Chickering,D.M. (1995) Learning Bayesian networks: The combination of knowledge and statistical data. *Mach. Learn.*, **20**, 197–243.
26. Kullback,S. and Leibler,R.A. (1951) On Information and Sufficiency. *Ann. Math. Stat*, **22**, 79–86.
27. Linhart,C., Halperin,Y. and Shamir,R. (2008) Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. *Genome Res.*, **18**, 1180–1189.
28. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
29. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (08, 2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
30. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (Jun, 2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, **316**, 1497–1502.
31. Wederell,E.D., Bilenky,M., Cullum,R., Thiessen,N., Dagpinar,M., Delaney,A., Varhol,R., Zhao,Y., Zeng,T., Bernier,B. *et al.* (2008) Global analysis of *in vivo* Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.*, **36**, 4549–4564.
32. Bradley,R.K., Li,X.-Y., Trapnell,C., Davidson,S., Pachter,L., Chu,H.C., Tonkin,L.A., Biggin,M.D. and Eisen,M.B. (03, 2010) Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila Species. *PLoS Biol.*, **8**, e1000343.
33. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (06, 2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem Cells. *Cell*, **133**, 1106–1117.
34. Corbo,J.C., Lawrence,K.A., Karlstetter,M., Myers,C.A., Abdelaziz,M., Dirkes,W., Weigelt,K., Seifert,M., Benes,V., Fritsche,L.G. *et al.* (2010) CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Res.*, **20**, 1512–1525.
35. Lin,Y.C., Jhunjhunwala,S., Benner,C., Heinz,S., Welinder,E., Mansson,R., Sigvardsson,M., Hagman,J., Espinoza,C.A., Dutkowski,J. *et al.* (2010) A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat. Immunol.*, **11**, 635–643.
36. He,A., Kong,S.W., Ma,Q. and Pu,W.T. (2011) Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc. Natl Acad. Sci. USA*, **108**, 5632–5637.
37. Heng,J.C.D., Feng,B., Han,J., Jiang,J., Kraus,P., Ng,J.H., Orlov,Y.L., Huss,M., Yang,L., Lufkin,T. *et al.* (2010) The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of Murine somatic cells to Pluripotent cells. *Cell Stem Cell*, **6**, 167–174.
38. Chen,L., Wu,G. and Ji,H. (2011) hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. *Bioinformatics*, **27**, 1447–1448.
39. Zhu,C., Byers,K.J., McCord,R.P., Shi,Z., Berger,M.F., Newburger,D.E., Saulrieta,K., Smith,Z., Shah,M.V., Radhakrishnan,M. *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.
40. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Res.*, **37(Suppl. 1)**, D77–D82.
41. Sandelin,A., Alkema,W., Engström,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
42. Hu,M., Yu,J., Taylor,J.M.G., Chinnaiyan,A.M. and Qin,Z.S. (2010) On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.*, **38**, 2154–2167.
43. ten Bosch,J.R., Benavides,J.A. and Cline,T.W. (2006) The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription. *Development*, **133**, 1967–1977.
44. Harrison,M.M., Li,X.Y., Kaplan,T., Botchan,M.R. and Eisen,M.B. (2011) Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-Zygotic transition. *PLoS Genet.*, **7**, e1002266.
45. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
46. Macindoe,I., Glockner,L., Vukasin,P., Stennard,F.A., Costa,M.W., Harvey,R.P., Mackay,J.P. and Sunde,M. (2009) Conformational stability and DNA binding specificity of the cardiac T-Box transcription factor Tbx20. *J. Mol. Biol.*, **389**, 606–618.
47. Sakabe,N.J., Aneas,I., Shen,T., Shokri,L., Park,S.Y., Bulyk,M.L., Evans,S.M. and Nobrega,M.A. (2012) Dual transcriptional activator and repressor roles of TBX20 regulate adult cardiac structure and function. *Hum. Mol. Genet.*, **21**, 2194–2204.
48. Kulakovskiy,I., Levitsky,V., Oshchepkov,D., Bryzgalov,L., Vorontsov,I. and Makeev,V. (2013) From binding motifs in ChIP-seq data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.*, **11**, 1340004.