Motif clustering with implications for transcription factor interactions

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

[1]**Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle, Germany**
[2]**German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany**
[3]**Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Quedlinburg, Germany**

ABSTRACT

High-throughput data, for instance ChIP-seq data, measure binding of transcription factors (TFs) or other proteins to DNA and have become a widespread data source for de-novo motif discovery. Often, several ChIP-seq data sets study the same TF under different conditions resulting in several, potentially redundant motifs, which demands for identification and clustering of similar motifs. Here, we propose a refined measure of motif similarity based on the correlation between score profiles on de Bruijn sequences. We demonstrate the utility of the proposed measure in benchmark studies on artificial motifs and motifs discovered from ENCODE ChIP-seq data. We use this measure to cluster motifs discovered from 757 different ENCODE ChIP-seq data sets for 166 TFs and RNA-polymerase II and III. Based on this clustering, we derive a TF interaction network that reflects many known TF-TF interactions, but also reveals novel putative interaction partners.

Keywords: motif, ChIP-seq, de Bruijn sequence, motif similarity, clustering, network

INTRODUCTION

Gene expression is regulated on the transcriptional level, amongst others, by transcription factors (TFs), which bind to genomic DNA and activate or repress the transcription of their target genes. Most transcription factors bind to DNA with some sequence specificity, which is typically described by sequence motifs. The development of high-throughput techniques lead to an ever-increasing number of experimental data sets that can be used for inferring binding motifs computationally. Techniques like ChIP-chip (Ren et al., 2000), ChIP-seq (Johnson et al., 2007), ChIP-exo (Rhee and Pugh, 2011), or ORGANIC (Kasinathan et al., 2014) allow for studying the *in-vivo* binding regions of a target TF on a genomic scale. One extensive resource of TF ChIP-seq data in human is the ENCODE project (The ENCODE Project Consortium, 2012).

The increasing amount of ChIP-seq data sets being generated also results in an increased redundancy on the motif level. The same TF is studied by different labs, in different cell types, at different developmental stages, and under different conditions. In addition, different approaches for *de-novo* motif discovery like MEME-ChIP (Machanick and Bailey, 2011), POSMO (Ma et al., 2012), ChIPMunk (Kulakovskiy et al., 2010), or Dimont (Grau et al., 2013) may be applied to infer motifs of the TF at hand. These aspects lead to some variation in the motifs obtained for a TF. While the primary motifs of sequence-specific TFs may be expected to show substantial similarity, secondary motifs or motifs originating from co-localization of binding sites may differ between cell types or approaches. Hence, this situation demands for clustering of similar motifs in a large set of inferred motifs based on some measure of motif similarity. In addition, such a similarity measure may help to identify the TF belonging to a motif. This is especially relevant for non-targeted approaches discovering motifs, e.g., from the promoters of co-expressed genes that are assumed to be regulated by a common but unknown TF.

Position weight matrix (PWM) models (Stormo et al., 1982; Staden, 1984) are still the most prevalent representation of TF binding motifs, but several approaches have been proposed recently that employ more complex motif models and may also capture dependencies between motif positions (Grau et al., 2013; Mordelet et al., 2013; Kulakovskiy et al., 2013; Eggeling et al., 2014).

Several measures of motif similarity have been proposed previously. Popular ones are column-based measures, which align two PWMs and then apply some similarity measure to each pair of

aligned PWM columns. The final similarity is then defined as the mean of these column-wise similarity values (Harbison et al., 2004; Mahony et al., 2007; Linhart et al., 2008). Column-wise similarity measures are, for instance, based on the Pearson correlation coefficient, the symmetric Kullkack-Leibler divergence, or the Euclidean distance between the PWM columns. Column-based measures are implemented in the popular tools Stamp (Mahony and Benos, 2007) and Tomtom (Gupta et al., 2007).

A more sophisticated measure for the similarity of position frequency matrices (PFMs, essentially non-normalized PWMs) based on the covariance of PFM hits is proposed by Pape et al. (2008). Kielbasa et al. (2005) propose a similarity measure, which is based on the Pearson correlation between the score profiles obtained from two motif models for a given input sequence. The authors consider random sequences with uniformly distributed bases for this purpose. In contrast to most other approaches, this similarity measure is not limited to PWM models but may be applied to motif models of arbitrary complexity that can compute a score for each sub-sequence under a sliding window.

Here, we propose a modification of the measure of Kielbasa et al. (2005) using a de Bruijn sequence of length 4^k containing all k-mers exactly once instead of a random sequence. De Bruijn sequences are cyclic and, hence, allow for a cyclic shift of score profiles to compensate for motif shifts.

We compare the performance of the proposed similarity measure to column based measures, Mosta (Pape et al., 2008), and the approach of Kielbasa et al. (2005) using artificial motifs as well as motifs discovered from ENCODE data. We apply the proposed measure to a large collection of motifs obtained by *de-novo* motif discovery using Dimont (Grau et al., 2013) on 757 ChIP-seq data sets for 166 TFs and RNA-polymerase II and III, clustering these motifs by similarity.

We further present and discuss selected clusters of motifs obtained by the proposed similarity measure. Finally, we use the result of motif clustering to build an interaction network of transcription factors based on motif co-occurrence.

METHODS

In this section, we define the measure of motif similarity, describe the benchmarking procedure, explain how motifs are clustered using the proposed similarity measure, describe how motif clustering is employed to build TF interaction networks, and introduce the data sets studied.

Motif similarity based on de Bruijn sequences

We denote a sequence of length L over the DNA alphabet by $\boldsymbol{x} = x_1 \ldots, x_L$ with $x_\ell \in \{A,C,G,T\}$, and a sub-sequence of length w starting at position ℓ in $\boldsymbol{x}$ by $\boldsymbol{x}_{\ell,w} = x_\ell \ldots x_{\ell+w-1}$. We further denote an arbitrary motif model by m, which may be applied to sequences of length w and returns a finite score $m(\boldsymbol{x})$ for each possible input sequence $\boldsymbol{x} \in \{A,C,G,T\}^w$. For instance, the score of a PWM model m_{PWM} with parameters $p_{\ell,a}$ corresponding to the probability of observing an $a \in \{A,C,G,T\}$ at position ℓ may be computed as

$$m_{\text{PWM}}(\boldsymbol{x}) = \sum_{\ell=1}^{w} \log\left(p_{\ell,x_\ell}\right). \tag{1}$$

We further use $s_{m,\ell} := m(\boldsymbol{x}_{\ell,w})$ as a shorthand for the score returned by m for the sub-sequence starting at position ℓ, and we define the *score profile* $\boldsymbol{s}_m = s_{m,1}, s_{m,2}, \ldots, s_{m,L-w+1}$ as the vector of all scores under a sliding window of length w in the test sequence $\boldsymbol{x}$. In analogy to Kielbasa et al. (2005), we then define the *correlation of score profiles* of two motif models m and m' as the Pearson correlation coefficient $\rho(\boldsymbol{s}_m, \boldsymbol{s}_{m'})$ of the corresponding two score profiles. We need to compensate for potentially different strand orientations of the two motifs m and m'. To this end, we compute a second score profile $\boldsymbol{s}_{m^r}$ with elements $s_{m^r,\ell} := m(rc(\boldsymbol{x}_{\ell,w}))$, where $rc(\boldsymbol{x}_{\ell,w})$ denotes the reverse complement of the sub-sequence $\boldsymbol{x}_{\ell,w}$ starting at position ℓ.

We extend the approach of Kielbasa et al. (2005) by using a de Bruijn sequence of length 4^k as test sequence $\boldsymbol{x}$. A de Bruijn sequence is defined as a cyclic sequence that contains all possible k-mers over the DNA alphabet exactly once. We compute $s_{m,\ell}$ for all sub-sequences of length w of the cyclic de Bruijn sequence. Hence, we always obtain a score profile $\boldsymbol{s}_m^+$ of length 4^k independent of the motif length w. The obtained score profile can itself be regarded as being cyclic. This facilitates compensation for shifts of one motif relative to the other, as a shift of motif m by i positions to the right (left) corresponds to a cyclic shift of $\boldsymbol{s}_m^+$ by i positions to the left (right), which we denote by $\boldsymbol{s}_{m\to i}^+$ ($\boldsymbol{s}_{m\to -i}^+$).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1302v1 | CC-BY 4.0 Open Access | rec: 13 Aug 2015, publ: 13 Aug 2015

We then define the similarity score of two motifs m and m' with a maximum shift of h as

$$\mathrm{sim}(m,m') = \max_{-h\le i\le h} \left\{ \max \left\{ \rho(\boldsymbol{s}^+_{m\to i}, \boldsymbol{s}^+_{m'}), \rho(\boldsymbol{s}^+_{m^r\to i}, \boldsymbol{s}^+_{m'}) \right\} \right\} \tag{2}$$

We permit all shift values h that result in an overlap of at least one third of the motif positions.

The proposed similarity measure entails a natural way of compensating for motif shifts by cyclic shifts of score profiles. In addition, it has the advantage that for motifs of length $w \le k$, the score profile over the de Bruijn sequence covers all w-mers with the same frequency of 4^{k-w} and, hence, fully and unbiasedly captures motif models of arbitrary complexity, for instance, any probability distribution $P(\boldsymbol{x})$ over the possible w-mers in $\{A,C,G,T\}^w$.

Using a de Bruijn sequence instead of a random sequence, hence, entails the advantage that shorter test sequences and score profiles are required to fully capture the probability distribution of a motif model. While a de Bruijn sequence of length $L = 4^w$ covers all w-mers by construction, a random sequence of the same length L covers on average only approximately $1 - \left(1 - \frac{1}{L}\right)^L \approx 1 - \frac{1}{e} \approx 63\,\%$ of all w-mers. The expectation of the length of a random sequence covering all w-mers at least once may be approximated by $E[L] \approx 4^w \cdot (\ln(4^w) + 0.577)$ (Blom et al., 1994), where the factor $\ln(4^w) + 0.577$ amounts to approximately 8.9 for $w = 6$ and 14.4 for $w = 10$.

Shorter de Bruijn sequences cover a subset of w-mers proportional to the sequence length. For instance, a de Bruijn sequence of length 4^{w-1} covers $1/4$-th of all w-mers. Different de Bruijn sequences exist for a given length and the exact subset of w-mers represented in a de Bruijn sequences shorter than 4^w depends on the generating algorithm.

Efficient algorithms for generating a de Bruijn sequence of length 4^k (Fredricksen and Maiorana, 1978; Ruskey, 2003) introduce undesired dependencies between positions at the left and right border if the motif length w becomes larger than k. For this reason, we generate the de Bruijn sequence $\boldsymbol{x}$ of length 4^k using an explicit de Bruijn graph with edges for each of the 4^k k-mers. In this graph, the de Bruijn sequence corresponds to an Eulerian cycle, where at each node, we choose the visiting order of outgoing edges randomly but with a fixed seed.

Motif clustering

For clustering a set of motifs $m_1, m_2, \ldots, m_N$, we compute all pairwise similarities $\mathrm{sim}(m_i, m_j)$. Most clustering methods are defined on distances or dissimilarities rather than similarities. We can easily convert the similarity proposed here to a dissimilarity as $\mathrm{dis}(m_i, m_j) := 1 - \mathrm{sim}(m_i, m_j)$, where $\mathrm{dis}(m_i, m_j) \in [0,2]$ and $\mathrm{dis}(m_i, m_i) = 0$.

We cluster the motifs using UPGMA (Sokal and Michener, 1958), i.e., agglomerative hierarchical clustering using average linkage, which yields a binary tree with motifs at the leaves and edge lengths representing dissimilarities. To obtain individual clusters, we cut this tree at height t, where each sub-tree corresponds to a cluster with an average pairwise similarity of its elements of at least $1 - t$.

When clustering PWMs, we additionally obtain for each inner node of the cluster tree a consensus PWM model by aligning all motifs using orientation and shift according to equation (2) recursively in the order induced by the tree structure. We compute the consensus PWM considering each column in the alignment separately and assuming a uniform distribution for missing motif positions due to shifts. We finally trim the consensus motif from its borders until we reach a position with an information content of at least 0.25 bit.

Data

For benchmarking purposes, we generate a set of 20 artificial motifs of length 10 bp by drawing columns of PWM models independently from a Dirichlet distribution with an equivalent sample size (Buntine, 1991) of 1. For each of the 20 motifs, we generate 40 random variations. Specifically, we (i) randomly draw a strand orientation, (ii) draw shifts in $[-4,4]$ from a uniform distribution and fill additional positions with a uniform distribution, and (iii) introduce noise by drawing 50 sequences from the corresponding distribution and re-estimating PWM parameters.

For all further analyses, we consider 757 ChIP-seq data sets from ENCODE (The ENCODE Project Consortium, 2012) for 166 TFs and RNA-polymerase II and III, which have been all ENCODE ChIP-seq data set with peaks from the ENCODE uniform peak calling pipeline available at the time of retrieval (10/2014). For each data set, we extract sequences of length 1,000 bp around all peak centers and annotate these with the corresponding peak statistics. We apply the de-novo motif discovery tool Dimont (Grau et al., 2013) to each of these data sets using a motif length of 20 bp and all remaining parameters set to their defaults, obtaining a variable number (including 0) of motifs represented as PWM models, which serve as the input motifs for the following steps.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1302v1 | CC-BY 4.0 Open Access | rec: 13 Aug 2015, publ: 13 Aug 2015

Benchmarking

In case of the artificial motifs, benchmarking is straightforward. Here, we know which sets of 40 motifs originate from the same initial PWM models and, hence, should be clustered together. Considering one of the motifs in one of these sets, we compute the similarity (i) to all other motifs originating from the same initial PWM and (ii) to the remaining, unrelated motifs. The corresponding similarity scores may be perceived as a classification score, where the scores of set (i) should receive larger scores than those in set (ii). We consider varying classification thresholds and measure the ability of a similarity measure to distinguish these two sets using the area under the precision-recall curve (AUC-PR) (Davis and Goadrich, 2006; Keilwagen et al., 2014). We choose AUC-PR instead of, e.g., the area under the ROC curve because of the skewed class ratio. We compute AUC-PR using the R-package PRROC (Grau et al., 2015).

We also benchmark the different approaches on the motifs obtained for the ENCODE data. In this case, there are different data sets from different labs or in different cell types for the same TF. We would expect that all data sets for the same TF share a common (e.g., primary) motif. Hence, for each pair of data sets, we compute the maximum similarity across all pairs of motifs discovered from these two data sets. When comparing data sets for different TFs, we adopt the same schema, where the largest similarity value represents the score of the "best false positive hit" between the different TFs. This procedure leads to pairwise similarity scores between all pairs of data sets. Following the same rationale as for the artificial motifs, we now quantify by AUC-PR how well a similarity measure can distinguish the data sets of identical from differing TFs. We limit this benchmark to those TFs with at least 10 data sets available and compute mean AUC-PR values and standard errors.

Building TF interaction networks

For each of the 166 TFs and RNA-polymerase II and III, the ENCODE data contain a number of experiments, and for each of the experiments, we obtain a number of motifs. We cluster all of these motifs by their dissimilarity and in the following regard motifs in a common cluster as similar and those in different clusters as dissimilar.

We consider two TFs T_i and T_j with N_i and N_j discovered motifs, respectively. If a substantial subset of these motifs is similar according to clustering, we regard these two TFs as interacting. We count for each of the K clusters and both TFs the numbers of motif occurrences $\boldsymbol{o}_i = (o_{i,1}, \ldots, o_{i,K})$ and $\boldsymbol{o}_j = (o_{j,1}, \ldots, o_{j,K})$. We then obtain the number $C_{i,j}$ of similar motif pairs of T_i and T_j as $C_{i,j} := \sum_{k=1}^{K} o_{i,k} o_{j,k}$, which is a measure for the degree of interaction between T_i and T_j. We assess the relevance of the obtained value $C_{i,j}$ by generating $10,000$ randomly permuted instances of the count vectors $\boldsymbol{o}_i$ and $\boldsymbol{o}_j$, computing the corresponding values $C'_{i,j}$, and counting the number $M_{i,j}$ of $C'_{i,j}$ values of at least $C_{i,j}$. We finally build an interaction network of TFs, in which the TFs correspond to nodes and edges between pairs of TFs are weighted with the corresponding $M_{i,j}$ values. Applying a threshold q on the edge weights, we omit edges with $M_{i,j} > q$ for clarity of a graphical network representation.

RESULTS AND DISCUSSION

In this section, we present the results of the benchmark studies, evaluate clustering of motifs discovered from ENCODE data, and analyze the derived TF interaction networks.

Benchmarks

We investigate if the conceptual advantage of using a de Bruijn sequence also has an influence on the accuracy of the two measures based on score profiles. To this end, we consider artificial motifs of length 10 using AUC-PR as performance measure (see Methods).

We first compare the score profile-based measures using de Bruijn sequences and random sequences on test sequences of different lengths in Figure 1(a). We find that, for short sequences, the similarity measure based on de Bruijn sequences indeed yields a greater accuracy than the measure based on a random test sequence. The difference between the two measures decreases with increasing sequence length and the accuracy of both approaches becomes highly similar for sequences of length 4^7 and above.

Longer or more complex motif models may model more specific binding to a small subset of sequences and, hence, may yield score profiles with more pronounced peaks. In such cases, the presence of a small subset of w-mers becomes important and the conceptual advantage of using de Bruijn sequences might become relevant for longer sequences as well.

In Figure 1(b), we compare the performance of the two similarity measures based on the correlation of score profiles using sequences of length 4^{10} with three column-based measures using as

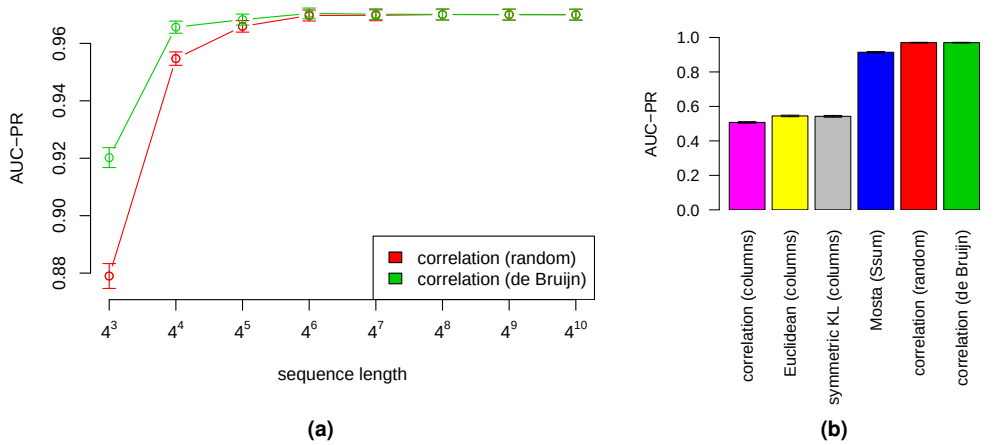


Figure 1. Comparison of performance on artificial motifs. In panel (a), we compare the measures based on correlation of score profiles using random sequences (red) and de Bruijn sequences (green) of different lengths. In panel (b) we consider a test sequence of length 4^{10} for these measures and additionally include column-based measures and Mosta into the comparison. Error bars indicate standard errors.

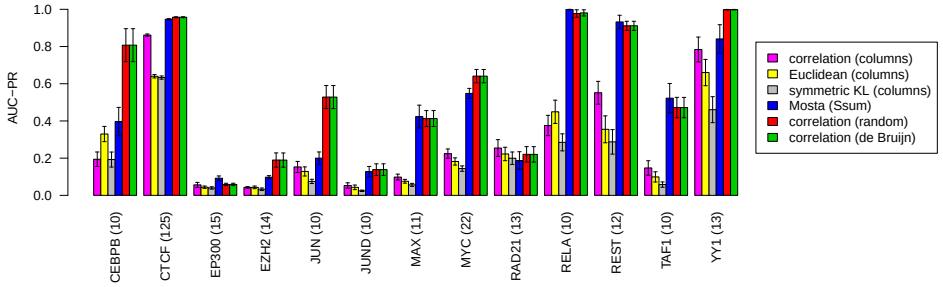


Figure 2. Comparison of similarity measures on motifs discovered from ENCODE ChIP-seq data sets for different TFs. Error bars indicate standard errors.

column-wise measures Pearson correlation, Euclidean distance, and the symmetric Kullback-Leibler divergence (Harbison et al., 2004; Mahony et al., 2007; Linhart et al., 2008), and with Mosta (Pape et al., 2008). We find that Mosta and the two measures based on the correlation of score profiles yield a substantially increased AUC-PR compared to the column-based measures.

One reason for this observation might be that the column-based measures are prone to false positive matches if two motifs share some similar bordering positions. In this case, shifted variants of these motifs might result in a high similarity, because additional, possibly dissimilar, positions are no longer aligned between the two motifs. However, this observations has limited implications for tools like Stamp (Mahony and Benos, 2007) or Tomtom (Gupta et al., 2007), because in these tools, the column-based measures are embedded in larger frameworks that also assess the significance of motif matches. The measures based on correlations of score profiles entail a conceptual advantage in this case, because relative shifts of the two motifs compared correspond to (cyclic) shifts of the corresponding score profiles. Hence, these measures always consider all motif positions, regardless of the current relative shift. On the artificial motifs, the two measures based on correlation of score profiles also yield a greater accuracy than Mosta.

We further compare these six similarity measures on data sets from the ENCODE project (Figure 2). Here, we assume that ChIP-seq data sets of the same TF share a similar motif, whereas no similar motifs are discovered in the data sets for other transcription factors (see Methods). Despite its limitation outlined below, a benchmark on real motifs might convey a more realistic picture of similarity measures under the scenario that we search a data base with a query motif of an unknown factor and aim at a hypothesis for the corresponding TF.

We exclude from the analysis the motifs discovered from RNA-polymerase II and III data sets, because these are expected to contain rather unspecific motifs, and we exclude all TFs with less than 10 data sets, because a low number of data sets is insufficient to compute reliable AUC-PR

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1302v1 | CC-BY 4.0 Open Access | rec: 13 Aug 2015, publ: 13 Aug 2015

values. In general, we find that the performance varies between the different TFs, which partly may be accounted for by different motif lengths and different levels of motif conservation. One further reason is indirect binding of TFs via other TFs (with specific binding motifs) to DNA. For instance, Rad21 is known to interact with CTCF (Wendt et al., 2008), and EP300 binds to CREB (Vo and Goodman, 2001), but both do not directly bind to genomic DNA. In addition, MAX and MYC bind to highly similar E-box motifs (Mordelet et al., 2013), and JUN and JUND both bind to the AP1 motif with consensus TGASTCA (Srivastava et al., 2013). However, all similarity measures should be affected by these limitations to a similar degree. Comparing the performance of the different similarity measures, we again observe an advantage of the score profile-bases measures and Mosta. In this case, Mosta and the score profile-based measures yield a similar accuracy.

In summary, we find that the similarity measure using de Bruijn sequences yields a higher accuracy than its analog using a random sequence for short sequences and works as well as state-of-the-art approaches like Mosta.

Similarity threshold

We further aim at a reasonable global threshold t on the similarity values, which may be used when searching a data base for motifs that are similar to a query motif. A reasonable threshold value should yield a high sensitivity and include a large fraction of the motifs of interest, but also yield a high precision and exclude a large fraction of false positives. A measure integrating both aspects is the F_1 measure, which is defined as the harmonic mean of sensitivity and precision. Hence, we follow a similar rationale as for the benchmarks and compute F_1 for different global classification thresholds t on the ENCODE data sets. We present boxplots of F_1 values for different values of t in Figure 3. We find a broad plateau of large median F_1 values for thresholds up to 0.3. The largest median F_1 values are observed for thresholds around 0.1, which may serve as a strict threshold focusing on precision. A threshold of 0.25 still yields a large median F_1 but balances between sensitivity and precision and, hence, may serve as a medium threshold.

Clustering of ENCODE motifs

We cluster the motifs discovered from the ENCODE data hierarchically by UPGMA using the similarity measure based on de Bruijn sequences. We obtain a partitioning of motifs into distinct groups by cutting the resulting binary tree at the medium threshold of $t = 0.25$.

We obtain 700 clusters for the 1541 motifs discovered from the ENCODE data sets. The majority of these clusters are singletons, resulting from (i) TFs with a single data set (82 singletons), (ii) motifs that are rather artifacts of de novo motif discovery, and (iii) motifs that are similar to those in other clusters, but are separated due to the choice of t, i.e., false negatives from the classification perspective. We also find 49 clusters with at least 5 motifs and 22 clusters with at least 10 motifs.

We present the cluster sizes of the five largest clusters with the corresponding consensus motifs in Figure 4. Despite its size, the consensus motif of the largest cluster clearly resembles the known CTCF motif. The consensus of the second cluster is similar to the known GABPA motif. It occurs in a large number of data sets targeting RNA-polymerase II, although this motif has not been described as RNA-polymerase II-related before. The consensus of the remaining three clusters show the known E-box motif of MYC, MAX, and MXI1 (Mordelet et al., 2013), the known GATA motif (Ko and

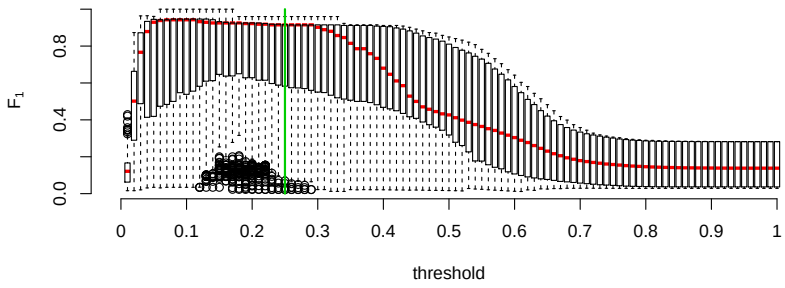


Figure 3. Choice of the threshold t. We plot boxplots of F_1 for different thresholds t on $\mathrm{dis}(m_i, m_j)$ highlighting the median in red. A threshold of $t = 0$ corresponds to a required similarity (i.e., Pearson correlation coefficient) of 1, while $t = 1$ corresponds to a required similarity of 0. The medium threshold of $t = 0.25$, which is also used for cutting the UPGMA cluster tree, is indicated by a green line.

TF	CTCF/RAD21	PolR2A	MYC/MAX/MXI1	GATA	JUN/FOS
#	147	97	56	38	37

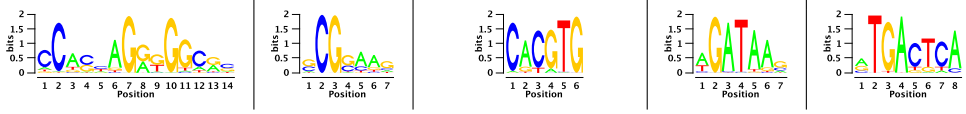


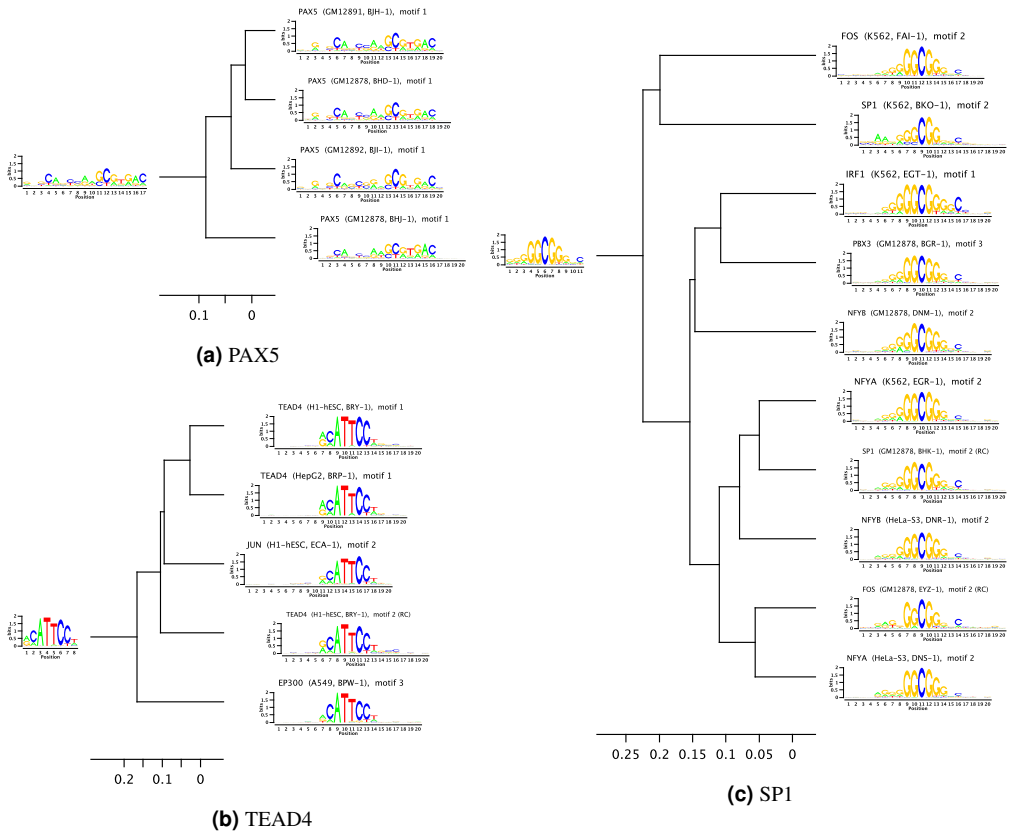
Figure 4. Cluster sizes and consensus motifs of the five largest clusters. We label each cluster with the TFs of the most prevalent motifs in each cluster.

Engel, 1993), and the known AP1 motif, which is bound by JUN/FOS hetero-dimers (Gustems et al., 2014).

In Figure 5, we show the cluster trees of three mid-sized clusters. The first cluster contains motifs discovered from four different ChIP-seq data sets for the TF PAX5. Notably, this motif is more degenerated than many other TF binding motifs. For this reason, we might have misjudged this motif as a motif discovery artifact, if it would have been present in only a single data set. However, the clustering reveals that a highly similar motif could be discovered form different data sets for PAX5, but not for other TFs, suggesting that this motif might be correct.

The second cluster contains three motifs for TEAD4 data sets and two similar motifs for a JUN and a EP300 data set. The motif is highly similar to the known TEAD4 motif (Benhaddou et al., 2012). The fact that the same motif has been discovered in two data sets for other TFs indicates a putative co-binding of TEAD4 with these TFs.

The third cluster shows that the known SP1 motif (Song et al., 2001) has been discovered not only in ChIP-seq data sets targeting SP1, but additionally in an even larger number of data sets for other TFs. In this case, we observe a co-clustering of motifs discovered from data sets of different TFs, for instance NFYA and PBX3, since both are related to each other by co-binding with SP1 as a common interaction partner.



(a) PAX5

(b) TEAD4

(c) SP1

Figure 5. Three representative cluster trees of mid-sized clusters. We show the consensus motif of each cluster at its root node. Motifs are labeled by the target TF, cell type and unique part of ENCODE experiment ID (in parentheses), and number of the discovered motif.

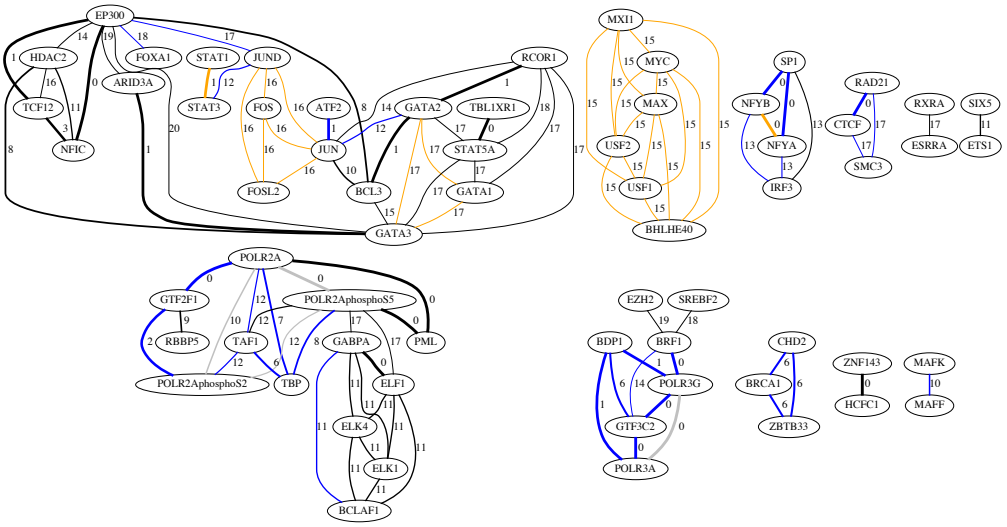


Figure 6. TF interaction network obtained by similarity of motifs discovered in ChIP-seq data sets for different TFs. Edge labels and widths indicate the number $M_{i,j}$ measuring the degree of interaction (see Methods), where lower values indicate greater confidence in an edge. Connected components comprising single nodes are omitted for clarity. Known interactions highlighted by colored edges: grey: essentially identical target; orange: known similar motif; blue: known interactions (Wang et al. (2012) and references in text).

TF interaction networks

Given distinct clusters of motifs obtained in the previous step, we build a TF interaction network based on the similarity of motifs discovered from ChIP-seq data sets of different TFs. Biological reasons for the occurrence of such similar motifs may be (i) different TFs binding to highly similar motifs (e.g., bHLH TFs) (ii) indirect binding to the DNA of one TF via another TF, (iii) a substantial co-localization of the binding sites of different TFs, which lead to the discovery of the motif of one TF in the ChIP-seq data set for another TF, or (iv) a third TF with binding sites co-localized with those of both of the two TFs. Here, we make the simplifying assumption that highly similar binding motifs also correspond to overlapping sets of binding sites in the genome, which might be further investigated by predictions in the original ChIP-seq positive regions in future studies.

We present the TF interaction network inferred from motif similarity in Figure 6. In this network, we observe several known interaction partners. For instance, RAD21 and SMC3 are known to interact with CTCF as part of the cohesin complex (Wendt et al., 2008); IRF3 is known to bind NFY motifs and SP1 is known to co-bind with NFY (Wang et al., 2012; Kheradpour and Kellis, 2014); and CHD2 and BRCA1 are known to indirectly bind to DNA via ZBTB33 (Wang et al., 2012). We also find connections between JUN, JUND, FOS, FOSL2, and ATF2 as part of a larger cluster, which are known to interact (Gustems et al., 2014). Another large connected component contains several TFs binding to E-boxes, namely MYC, MAX, MXI1, USF1, USF2, and BHLHE40 (Mordelet et al., 2013; Foley and Sidow, 2013).

We also observe a connection of RNA-polymerase II to several known basal TFs, namely GTF2F1, TAF1, TBP, which are involved in RNA-polymerase II-dependent transcription (Ruppert et al., 1993; Wang et al., 2012). Notably, RNA-polymerase III is connected to a completely different set of basal TFs, some of which are known to be important for RNA-polymerase III-dependent transcription (Liao et al., 2003).

These observations support the TF interactions derived from the occurrence of similar motifs and indicate that some of the less well-known interactions like those of RBBP5 and GTF2F1, ARID3A and EP300, TBL1XR1 and STAT5, or NFIC and TCF12 are promising candidates of genuinely interacting TF partners.

Conclusions

We present a novel measure of motif similarity that is based on the correlation of score profiles on de Bruijn sequences. In contrast to most existing measures, the proposed measure can be applied to arbitrary motif models. In benchmark studies, we show that this measure yields an accuracy that is comparable to state-of-the-art approaches. We use the proposed measure to cluster a large collection

of motifs inferred from ENCODE ChIP-seq data. Building an interaction network of TFs based on motif similarity, we find connections between several known interaction partners, but also discover novel putative interaction partners.

We provide a web-application at http://galaxy.informatik.uni-halle.de that searches for motifs similar to a query motif based on the proposed similarity measure. The query motif may be supplied as a PWM or PFM, as a set of aligned binding sites, or as a score profile. All motifs and motif clusters are also available for download in a text format at http://www.jstacs.de/index.php/DBcorrDB.

ACKNOWLEDGMENTS

We thank Yvonne Pöschl for valuable discussions and Thomas Berner for technical support.

REFERENCES

Benhaddou, A., Keime, C., Ye, T., Morlon, A., Michel, I., Jost, B., Mengus, G., and Davidson, I. (2012). Transcription factor TEAD4 regulates expression of Myogenin and the unfolded protein response genes during C2C12 cell differentiation. *Cell Death Differ*, 19(2):220–231.

Blom, G., Holst, L., and Sandell, D. (1994). *Problems and Snapshots from the World of Probability*. Springer-Verlag, New York, 1st edition.

Buntine, W. L. (1991). Theory refinement of Bayesian networks. In *Uncertainty in Artificial Intelligence*, pages 52–62. Morgan Kaufmann.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, New York, NY, USA. ACM.

Eggeling, R., Gohr, A., Keilwagen, J., Mohr, M., Posch, S., Smith, A. D., and Grosse, I. (2014). On the value of intra-motif dependencies of human insulator protein CTCF. *PLoS ONE*, 9(1):e85629.

Foley, J. and Sidow, A. (2013). Transcription-factor occupancy at HOT regions quantitatively predicts RNA polymerase recruitment in five human cell lines. *BMC Genomics*, 14(1):720.

Fredricksen, H. and Maiorana, J. (1978). Necklaces of beads in k colors and k-ary de Bruijn sequences. *Discrete Mathematics*, 23:207–210.

Grau, J., Grosse, I., and Keilwagen, J. (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*.

Grau, J., Posch, S., Grosse, I., and Keilwagen, J. (2013). A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Research*, 41(21):e197.

Gupta, S., Stamatoyannopoulos, J., Bailey, T., and Noble, W. (2007). Quantifying similarity between motifs. *Genome Biology*, 8(2):R24.

Gustems, M., Woellmer, A., Rothbauer, U., Eck, S. H., Wieland, T., Lutter, D., and Hammerschmidt, W. (2014). c-Jun/c-Fos heterodimers regulate cellular genes via a newly identified class of methylated DNA sequence motifs. *Nucleic Acids Research*, 42(5):3059–3072.

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., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., and Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502.

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 Meth*, 11(2):203–209.

Keilwagen, J., Grosse, I., and Grau, J. (2014). Area under precision-recall curves for weighted and unweighted data. *PLoS ONE*, 9(3):e92209.

Kheradpour, P. and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE tf binding experiments. *Nucleic Acids Research*, 42(5):2976–2987.

Kielbasa, S., Gonze, D., and Herzel, H. (2005). Measuring similarities between transcription factor binding sites. *BMC Bioinformatics*, 6(1):237.

Ko, L. J. and Engel, J. D. (1993). DNA-binding specificities of the GATA transcription factor family. *Molecular and Cellular Biology*, 13(7):4011–4022.

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. *Journal of Bioinformatics and Computational Biology*, 11(01):1340004.

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(20):2622–2623.

Liao, Y., Willis, I. M., and Moir, R. D. (2003). The Brf1 and Bdp1 subunits of transcription factor TFIIIB bind to overlapping sites in the tetratricopeptide repeats of Tfc4. *Journal of Biological Chemistry*, 278(45):44467–44474.

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 Research*, 18(7):1180–1189.

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 Research*, 40(7):e50.

Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–1697.

Mahony, S., Auron, P. E., and Benos, P. V. (2007). DNA familial binding profiles made easy: Comparison of various motif alignment and clustering strategies. *PLoS Comput Biol*, 3(3):e61.

Mahony, S. and Benos, P. V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research*, 35(suppl 2):W253–W258.

Mordelet, F., Horton, J., Hartemink, A. J., Engelhardt, B. E., and Gordân, R. (2013). Stability selection for regression-based models of transcription factor–DNA binding specificity. *Bioinformatics*, 29(13):i117–i125.

Pape, U. J., Rahmann, S., and Vingron, M. (2008). Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics*, 24(3):350–357.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309.

Rhee, H. S. and Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419.

Ruppert, S., Wang, E. H., and Tjian, R. (1993). Cloning and expression of human TAFII250: a TBP-associated factor implicated in cell-cycle regulation. *Nature*, 362(6416):175–179.

Ruskey, F. (2003). *Combinatorial generation*. Working Version (1j-CSC 425/520). Department of Computer Science, University of Victoria, Victoria, Canada.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438.

Song, J., Ugai, H., Ogawa, K., Wang, Y., Sarai, A., Obata, Y., Kanazawa, I., Sun, K., Itakura, K., and Yokoyama, K. K. (2001). Two consecutive zinc fingers in Sp1 and in MAZ are essential for interactions with cis-elements. *Journal of Biological Chemistry*, 276(32):30429–30434.

Srivastava, P., Hull, R., Behmoaras, J., Petretto, E., and Aitman, T. (2013). JunD/AP1 regulatory network analysis during macrophage activation in a rat model of crescentic glomerulonephritis. *BMC Systems Biology*, 7(1):93.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, 12:505–519.

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 Research*, 10(9):2997–3010.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Vo, N. and Goodman, R. H. (2001). CREB-binding protein and p300 in transcriptional regulation. *Journal of Biological Chemistry*, 276(17):13505–13508.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M., and Weng, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 22(9):1798–1812.

Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., Yahata, K., Imamoto, F., Aburatani, H., Nakao, M., Imamoto, N., Maeshima, K., Shirahige, K., and Peters, J.-M. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, 451(7180):796–801.