Phylogenetics

Unrealistic phylogenetic trees may improve phylogenetic footprinting

Martin Nettling¹,* Hendrik Treutler², Jesus Cerquides³ and Ivo Grosse¹,⁴

¹Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle, Germany, ²Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Halle, Germany, ³Institut d’Investigació en Intel·ligència Artificial, I3IA-CSIC, Campus UAB, Cerdanyola, Spain and ⁴German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: The computational investigation of DNA binding motifs from binding sites is one of the classic tasks in bioinformatics and a prerequisite for understanding gene regulation as a whole. Due to the development of sequencing technologies and the increasing number of available genomes, approaches based on phylogenetic footprinting become increasingly attractive. Phylogenetic footprinting requires phylogenetic trees with attached substitution probabilities for quantifying the evolution of binding sites, but these trees and substitution probabilities are typically not known and cannot be estimated easily.

Results: Here, we investigate the influence of phylogenetic trees with different substitution probabilities on the classification performance of phylogenetic footprinting using synthetic and real data. For synthetic data we find that the classification performance is highest when the substitution probability used for phylogenetic footprinting is similar to that used for data generation. For real data, however, we typically find that the classification performance of phylogenetic footprinting surprisingly increases with increasing substitution probabilities and is often highest for unrealistically high substitution probabilities close to one. This finding suggests that choosing realistic model assumptions might not always yield optimal predictions in general and that choosing unrealistically high substitution probabilities close to one might actually improve the classification performance of phylogenetic footprinting.

Availability and Implementation: The proposed PF is implemented in JAVA and can be downloaded from https://github.com/mgledi/PhyFoo

Contact: martin.nettling@informatik.uni-halle.de

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Gene regulation is a highly complex process in nature based on several sub-processes such as transcriptional regulation including DNA methylation (Smith and Meissner, 2013), histone modifications (Tessarz and Kouzarides, 2014) and promoter escaping (Sainsbury et al., 2015) as well as post-transcriptional regulation including modulated mRNA decay (Schoenberg and Maquat, 2012), siRNA interference (de Fougerolles et al., 2007; Tam et al., 2008) and alternative splicing (Luco et al., 2010; Sultan et al., 2008). One important step in this complex process is the regulation of transcriptional initiation by the interaction of transcription factors (TFs) with their binding sites (Hobert, 2008; Voss and Hager, 2014). Hence, identifying transcription factor binding sites (TFBSs) and inferring their binding motifs is a prerequisite in modern
FootPrinter

Examples of highly popular tools for phylogenetic footprinting are able genomes of different species and due to increasing computational resources, approaches for de-novo motif discovery based on phylogenetic footprinting have become increasingly attractive. Examples of highly popular tools for phylogenetic footprinting are FootPrinter (Blanchette and Tompa, 2003), PhyME (Sinha et al., 2004), MONKEY (Moses et al., 2004a), PhyloGibbs (Siddharthan et al., 2005), Phylogenetic Gibbs Sampler (Newberg et al., 2007), PhyloGibbs-MP (Siddharthan, 2008) and MotEvo (Arnold et al., 2012). Supplementary Table S1 provides a comparison of these tools regarding the used evolutionary model, sequence model and learning principle.

One prerequisite for most phylogenetic footprinting approaches are multiple sequence alignments (MSAs) of upstream regions of orthologous genes of multiple not too closely related species (Anisimova et al., 2013). These MSAs capture phylogenetic information, and the key idea of using these MSAs as starting point for phylogenetic footprinting results from the observations that (i) functional TFBSs are phylogenetically conserved and (ii) phylogenetically conserved TFBSs are aligned in MSAs. Examples of highly popular tools for aligning non-coding genomic regions are T-Coffee (Notredame et al., 2000), WebPRANK (Löytynoja and Goldman, 2010) and MAFFT (Katoh and Standley, 2013).

Phylogenetic footprinting improves the de-novo motif discovery by incorporating phylogenetic dependencies within the MSA in contrast to approaches based on sequences from only one species. Substitution models of DNA sequence evolution such as the F81 model (Felsenstein, 1981) have been adapted to model the evolution of TFBSs in a position-specific manner, and it has been shown that these adapted models, which we call phylogenetic footprinting models (PFMs) for brevity, can detect TFBSs more accurately than models that neglect phylogenetic dependencies (Clark et al., 2007; Gertz et al., 2006; Hardison and Taylor, 2012; Hawkins et al., 2009; Moses et al., 2004a; Nettling et al., 2017).

One fundamental prerequisite for phylogenetic footprinting is a phylogenetic tree including substitution probabilities attached to each of its branches, and choosing an appropriate phylogenetic tree and appropriate substitution probabilities is pivotal for the classification performance of phylogenetic footprinting (Kc and Livesay, 2011). However, estimating substitution probabilities within TFBSs is substantially harder than estimating them e.g. in protein-coding regions for at least two reasons:

First, the positions of TFBSs are unknown when performing phylogenetic footprinting, whereas the positions of protein-coding regions are known when estimating substitution probabilities there. Second, protein-coding regions are much longer than TFBSs, so one can use a much larger number of bases for estimating substitution probabilities for protein-coding regions than for TFBSs.

Estimating substitution probabilities within TFBSs is challenging, but several valuable studies have been performed in this direction (Doniger and Fay, 2007; Pollard et al., 2010; Schaalke et al., 2015; Tuğrul et al., 2015). For example, studies on synthetic data have indicated that small substitution probabilities in the motif and moderate substitution probabilities in the flanking sequences can be preferable for motif recognition (Sinha et al., 2004), and studies on different yeast species have confirmed these findings and shown that the likelihood of the Jukes-Cantor model (Jukes and Cantor, 1969) increases relative to a thymine background (‘polyT’) for small substitution probabilities in the motif and moderate substitution probabilities in the flanking sequences (Moses et al., 2004b).

These and similar findings, however, have not lead to a robust approach of estimating substitution probabilities within TFBSs prior to or as part of phylogenetic footprinting, so the substitution probabilities are often simply taken from the literature or guessed, and their influence on the classification performance of phylogenetic footprinting has not yet been studied systematically.

Here, we study this influence based on a synthetic dataset and five real datasets of the TFs CTCF, GABP, NRSF, SRF and STAT1.

Specifically, we describe the PFM, the datasets, the tested phylogenetic trees, the performance measure, and implementation details in section Methods, and we study the classification performance of phylogenetic footprinting as a function of the substitution rate for synthetic and real datasets, compare the results to those of phylogenetic footprinting based on expert trees from the literature, and discuss the findings in the context of several factors that affect the evolution of TFBSs in sections 3 and 4.

2 Materials and methods

In this section we describe (i) the used notation and the likelihood calculation of the PFM, (ii) the investigated datasets, (iii) the performance measure, (iv) the systematic investigation of phylogenetic trees and (v) the implementation of the PFMs.

2.1 Phylogenetic footprinting model

2.1.1 Notation

Our data contains \( N \) alignments, with each alignment containing \( O \) sequences (one per observed species) of length \( L_0 \).

Our phylogenetic model incorporates the existence of \( H \) additional hidden species, that is, species for which we cannot observe their sequences. Both hidden and observed species form a tree. Thus, for each species \( k \) but the root, \( p_a(k) \) denotes the ancestor of species \( k \) in the tree. The root species is noted \( r \).

Our probabilistic model contains a random variable \( S_{n,k}^u \) for each nucleotide \( 1 \leq u \leq L_n \) of each species \( 1 \leq k \leq O + H \) of each alignment \( 1 \leq n \leq N \). These random variables take values in the set of bases \( A = \{ A, C, G, T \} \). We note \( p_a(S_{n,k}^u) \) the \( u \)-th nucleotide in the \( r \)-th alignment of species \( p_a(k) \) (the ancestor of \( k \)). By definition, the root has no ancestor and hence \( p_a(S_{n,k}^u) = G \). We also refer to nucleotide \( S_{n,k}^u \) as \( A_{n,k}^u \) when species \( k \) is observed, and as \( Y_{n,k}^u \) when species \( k \) is hidden. Furthermore we note by \( Y_{n,k}^u \) (respectively \( S_{n,k}^u \)) the set containing each random variable \( Y_{n,k}^u \) (respectively \( S_{n,k}^u \)), with \( O + 1 \leq k \leq O + H \) and \( Y_n \) the set containing every random variable in \( Y_{n,k}^u \) with \( 1 \leq u \leq L_n \).

An alignment \( A_n \) may or may not contain a TFBS. This is encoded in variable \( M_n \) with \( M_n^a \) indicating that alignment \( A_n \) does not contain a motif and \( M_n^b \) indicating that alignment \( A_n \) does contain a motif.

2.1.2 Likelihood

The probability that the alignment \( A_n \) is generated by the PFM can be written as

\[
p(A_n | \theta) = p(A_n | M_n^b, \theta) \times p(M_n^b | \theta) + p(A_n | M_n^a, \theta) \times p(M_n^a | \theta)
\]

with variable \( M_n \) taking a Bernoulli distribution and \( \theta \) denoting model parameters, namely the topology of the phylogenetic tree, the substitution probabilities and the evolutionary model with its
stationary probabilities for the flanking regions as well as the TFBS regions.

We need to specify the probability for non-motif-bearing \( p(A_n | M^n_w, \theta) \) and for motif-bearing alignments \( p(A_n | M^n_w, \theta) \). For reasons of clarity we omit \( \theta \) in the following.

### 2.1.3 Likelihood of a non-motif-bearing alignment

The probability that alignment \( A_n \) is generated by the PFM as a non-motif bearing alignment is

\[
p(A_n | M^n_w) = \sum_{Y_n} p(A_n | Y_n, M^n_w).
\]

We assume that each single nucleotide alignment is independent of any other nucleotide alignment given \( \theta \) and \( M^n_w \). Furthermore, we assume that in each nucleotide alignment, the species satisfy the conditional independencies encoded by the phylogenetic tree. Thus,

\[
p(A_n | M^n_w) = \prod_{k=1}^{\nu_n} \sum_{\gamma_k} p(S^n_k | M^n_w)
\]

where

\[
p(S^n_k = a | pa(S^n_k) = b, M^n_w) = \begin{cases} \gamma_k \times \pi^n_a & \text{if } k = r \\ \gamma_k \times \pi^n_a + (1 - \gamma_k)\delta_{a,b} & \text{if } k \neq r \end{cases}
\]

generic function of \( \gamma_k \) model, where the base distribution of each position of the background sequence is denoted by \( \pi^n_0 \), the probability of a nucleotide \( a \) in the background sequence is denoted by \( \pi^n_a \), and the substitution probability from the ancestor species to species \( k \) is denoted by \( \gamma_k \). For more realistic phylogenetic models \( \gamma_k \) might also depend on specific nucleotide transitions.

### 2.1.4 Likelihood of a motif-bearing alignment

The probability that alignment \( A_n \) is generated by the PFM as a motif bearing alignment is

\[
p(A_n | M^n_w) = \sum_{\ell_n = 1}^{\ell_n = W - 1} \sum_{\epsilon_n = 1}^{\ell_n} p(A_n, Y_n, \ell_n | M^n_w).
\]

Where \( W \) is the length of the TFBS and \( \ell_n \) is the position of the TFBS in alignment \( A_n \). Since single nucleotide alignments are assumed independent and considering the conditional independencies in the phylogenetic tree we have

\[
p(A_n | M^n_w) = \sum_{\ell_n = 1}^{\ell_n = W - 1} \sum_{\epsilon_n = 1}^{\ell_n} p(A_n, Y_n, \ell_n | M^n_w). \tag{5}
\]

with

\[
p(S^n_k | \epsilon_n, M^n_w) = \prod_{k=1}^{\nu_n} p(S^n_k | pa(S^n_k), \epsilon_n, M^n_w) \quad \text{and} \quad p(S^n_k | pa(S^n_k), \epsilon_n, M^n_w) = \begin{cases} \gamma_k \times \pi^n_a & \text{if } k = r \\ \gamma_k \times \pi^n_a + (1 - \gamma_k)\delta_{a,b} & \text{if } k \neq r \end{cases}
\]

As for the non-motif-bearing alignment, the base distribution of each position of the background sequence is denoted by \( \pi^n_0 \) and the probability of a nucleotide \( a \) in the background sequence is denoted by \( \pi^n_a \). The base distributions of a motif sequence of length \( W \) are denoted by \( \pi^n_a \) with \( w \in \{1, \ldots, W\} \) and the probability of a nucleotide \( a \) at position \( w \) in a motif sequence is denoted by \( \pi^n_a \). The substitution probability from the ancestor species to species \( k \) is denoted by \( \gamma_k \).

Finally we assume motifs to be uniformly distributed, thus having that \( p(\ell_n | M^n_w) = \frac{1}{W-1} \), which completes the specification of the likelihood function.

### 2.2 Data

#### 2.2.1 Real data

The data used in this work originate from human ChIP-Seq data of the five Tfs CTCF, GABP, NRSF, SRF and STAT1 Joho et al. (2008); Valouev et al. (2008) and gapped alignments of the ChIP-Seq target regions from human with orthologous regions from monkey, cow, dog and horse. The original data provided by Arnold et al. (2012) consist of 900 gapped alignments for each of the five Tfs. Each gapped alignment consists of sequences from six species. Since gapped alignments have a higher risk of showing mathematical side effects, we process them to derive ungapped alignments following three steps: (i) We remove the species that causes the highest number of gaps in all alignments. Accordingly, we remove sequences from opossum and keep orthologous regions from human, monkey, cow, dog and horse. (ii) In each alignment, we remove all alignment columns that contain at least one gap. (iii) We remove all alignments that are shorter than 21 bp, which is the length of the longest TFBS motif (NRSF) in the presented studies. Supplementary Table S2 shows details about the resulting datasets. All datasets are available as Supplementary Material.

#### 2.2.2 Synthetic data

The synthetic dataset used in this work is generated using the PFM specified in section 2.1 with a star topology.

A negative set of 1000 non-motif-bearing alignments each of length \( L = 300 \) is generated. Each non-motif bearing alignment is generated in two steps as follows. (i) Sample the primordial sequence. For each position \( u \in [1, L] \) of the sequence, sample a nucleotide from the uniform distribution \( \pi^n_0 \). (ii) For each of the descent species \( o \in \{1, \ldots, 5\} \), sample a mutated sequence given the primordial sequence position-wise. For each position \( u \in [1, L] \), apply the F81 Felsenstein (1981) mutation model with the equilibrium distribution \( \pi^n_0 \) and substitution probability \( \gamma = 0.2 \) to the nucleotide of the primordial sequence at position \( u \).

A positive set of 750 motif-bearing alignments each of length \( L = 300 \) is generated. Each motif-bearing alignment is generated as follows:

(i) Sample the primordial sequence given a TFBS length of \( W = 15 \).

(a) Sample the start position \( \ell \in [1, L - W + 1] \) of the TFBS from the uniform distribution.

(b) For each position \( u \in [1, \ell - 1] \) and \( u \in [\ell + W, L] \) of the flanking position, we sample the nucleotide at position \( u \) from the uniform distribution \( \pi^n_0 \). For each position \( u \in [\ell, \ell + W - 1] \) of the TFBS, we sample the nucleotide at position \( u \) from the distribution \( \pi^n_{\ell - \ell + 1} \). The distribution \( \pi^n_{\ell - \ell + 1} \) is uniformly drawn from the simplex.

(ii) For each of the descent species \( o \in \{1, \ldots, 5\} \), sample a mutated sequence given the primordial sequence position-wise.

(a) For each position \( u \in [1, \ell - 1] \) and \( u \in [\ell + W, L] \) of the flanking sequence, apply the F81 mutation model with the equilibrium distribution \( \pi^n_0 \) and substitution probability \( \gamma = 0.2 \) to the nucleotide of the primordial sequence at position \( u \).

(b) For each position \( u \in [\ell, \ell + W - 1] \) of the TFBS, apply the F81 mutation model with the equilibrium distribution \( \pi^n_{\ell - \ell + 1} \) and substitution probability \( \gamma = 0.2 \) to the nucleotide of the primordial sequence at position \( u \).
2.3 Phylogenetic trees
To systematically investigate the influence of different phylogenetic trees on classification performance and hence on motif prediction, we introduce two simplifications. First, the underlying phylogenetic tree is a star topology implying that all species have one common ancestor. Second, all branches in the star topology have the same length, i.e. the probability that a base in the primordial sequence is replaced by a new base in a descendant sequence is the same for all sequences.

Now, it is possible to systematically vary the substitution probabilities \( \gamma = \{0.05, 0.1, \ldots, 1.0\} \), where \( \gamma \) is inversely proportional to the phylogenetic relatedness. Small \( \gamma \) encode close phylogenetic relations and large \( \gamma \) encode distant phylogenetic relations. Especially, \( \gamma = 1.0 \) implies that the species are phylogenetically unrelated, i.e. the sequences of each alignment are statistically independent.

2.4 Classification performance
We evaluate all PFMs by a stratified repeated random sub-sampling validation by estimating all PFMs from a training set and measuring classification performance on a test set as follows.

In step 1, we generate two training sets and two disjoint test sets for each of the five TFs as follows. We randomly select 200 alignments from the set of alignments of a particular TF as positive training set, leaving the remaining alignments as positive test set. We perform a base shuffling on the positive set of alignments of the same TF to get a negative set of alignments. We randomly select 200 alignments from this set of alignments as negative training set and leave the remaining alignments as negative test set.

In step 2, we train a foreground model on the positive training set and a background model on the negative training set by expectation maximization (Lawrence and Reilly, 1990) using a numerical optimization procedure in the maximization step. We restart the expectation maximization algorithm, which is deterministic for a given dataset and a given initialization, 20 times with different initializations and choose the foreground model and the background model with the maximum likelihood on the positive training data and the negative training data, respectively, for classification. We use a likelihood-ratio classifier of the two chosen foreground and background models, apply this classifier to the disjoint positive and negative test sets, and calculate the area under the receiver operating characteristics curve and the area under the precision recall curve as measures of classification performance.

We repeat both steps 100 times and determine (i) the mean area under the receiver operating characteristic curve and its standard error and (ii) the mean area under the precision recall curve and its standard error.

2.5 Implementation
In order to investigate the influence of different phylogenetic trees in a fair and detailed way, we implement the proposed PFM based on the freely available Java Framework Jstacs (Grau et al., 2012). Among others, Jstacs provides ready-to-use sequence models for reuse, numerical and non-numerical optimization procedures for model estimation, serialization of models and methods for the statistical evaluation of results. In contrast to existing tools which are typically focused on application, using Jstacs we are able to compare different PFMs in a detailed way by extracting mandatory information about the inferred models and the predicted TFBSs.

Algorithm 1 shows the pseudocode for inferring a PFM from a set of alignments. The implementation of the proposed PFM is available at https://github.com/mgledi/PhyFoo/.

3 Results
In this section, we investigate the classification performance of the PFM specified in section 2.1 as function of the substitution probability for a synthetic dataset and five real datasets. The synthetic dataset is generated using the PFM described in section 2.2. The five real datasets originate from human ChIP-Seq experiments of the five TFs CTCF, GABP, NRSF, SRF and STAT1 and MSAs of the predicted target regions with orthologous regions from monkey, cow, dog and horse as described in section 2.2.

In section 2.1.1, we study the likelihood of the popular PFM specified in section 2.2 as a function of the substitution probability for the synthetic dataset and the real dataset of TF CTCF. In section 2.1.2, we study the classification performance of the PFM as a function of the substitution probability for the same datasets. In section 2.1.3, we perform the studies of subsections 1 and 2 for the four datasets of the TFs GABP, NRSF, SRF and STAT1. In section 2.1.4, we study the classification performance of the PFM based on three selected phylogenetic trees for all five datasets of the TFs CTCF, GABP, NRSF, SRF and STAT1.

3.1 Likelihood on synthetic and real data
First, we test the implemented expectation maximization algorithm for the PFM specified in section 2.1 and summarized in Algorithm 1 by applying it to synthetic data generated with a substitution probability of 0.2 as described in section 2.2 and to real data of TF CTCF. In both cases, we vary the substitution probability \( \gamma \) of the PFMs from 0.05 to 1.0 with increments of 0.05.

In case of synthetic data, we expect the best fit of the PFMs and thus the highest likelihood when the substitution probability \( \gamma \) of the PFMs is close to the substitution probability of 0.2 used for data generation. In case of real data of TF CTCF, we expect the best fit of the PFMs and thus the highest likelihood when the substitution...
probability \( \gamma \) of the PFMs is in the range of \( 0.1 \leq \gamma \leq 0.4 \) according to \textit{Getz et al.} (2006).

Figure 1a shows the likelihood as a function of the substitution probability \( \gamma \) ranging from 0.05 to 1.0 with increments of 0.05 for synthetic data, and we observe the expected function with a maximum at the substitution probability of \( \gamma = 0.2 \), which is equal to the substitution probability used for data generation. Figure 1b shows the likelihood as a function of the substitution probability \( \gamma \) for real data of TF CTCF, and we again observe the expected function with a maximum at the substitution probability of \( \gamma = 0.2 \), which is a reasonable value and in the range of \( 0.1 \leq \gamma \leq 0.4 \) suggested by \textit{Getz et al.} (2006).

These findings indicate that the applied PFM and the applied maximum-likelihood principle are capable of identifying reasonable substitution probabilities for synthetic and real data of TF CTCF, where reasonable substitution probabilities mean substitution probabilities close to those used for data generation in case of synthetic data and in the range suggested by experts for real data of TF CTCF.

### 3.2 Classification performance on synthetic and real data

Second, we study the classification performance of the PFMs by the method described in section 2.3 on the same two datasets. We again vary \( \gamma \) from 0.05 to 1.0 with increments of 0.05 and compute the classification performance as a function of \( \gamma \) as described in section 2.4.

In case of both synthetic and real data, we expect that the classification performance looks qualitatively similar to the likelihood as a function of \( \gamma \); i.e. we expect that the classification performance is highest for \( \gamma \) close to 0.2 for synthetic data and in the range of \( 0.1 \leq \gamma \leq 0.4 \) for real data of TF CTCF.

Figure 2a shows the classification performance as a function of \( \gamma \) for synthetic data, and we observe the expected function with a maximum at \( \gamma = 0.2 \), which is equal to the substitution probability used for data generation and equal to the location of the maximum of the likelihood. These results are in agreement with those of \textit{Sinha et al.} (2004) who additionally find that an underestimation of the true substitution probability leads to a more severe degradation of the classification performance than an overestimation of equal magnitude.

Figure 2b shows the classification performance as a function of \( \gamma \) for real data of TF CTCF, but here we observe a function that is different from the expected function, different from the function observed for synthetic data, and different from the likelihood function of Figure 1b. Specifically, we observe that the maximum is achieved for an unrealistically high value of \( \gamma = 1.0 \), which is clearly outside of the range of substitution probabilities of \( 0.1 \leq \gamma \leq 0.4 \) suggested by \textit{Getz et al.} (2006) and much greater than the value of \( \gamma = 0.2 \) at which the maximum of the likelihood is located.

This observation is surprising because a substitution probability of \( \gamma = 1.0 \) corresponds to a PFM that assumes the orthologous sequences in the MSAs be statistically independent, i.e. phylogenetically unrelated. It indicates that choosing a realistic substitution probability in the range of \( 0.1 \leq \gamma \leq 0.4 \) might lead to an inferior classification performance of phylogenetic footprinting compared to choosing an unrealistically substitution probability of \( \gamma = 1.0 \).

### 3.3 Classification performance and likelihood on four additional real datasets

Third, we study if the phenomenon that the maximum classification performance is achieved for an unrealistically high value of \( \gamma \) is specific for TF CTCF or possibly also present in other TFs. Hence, we perform the studies of sections 2.2.1 and 2.2.2 for four additional ChIP-Seq datasets of TFs GABP, NRSF, SRF and STAT1.

Figure 3a–d shows the four classification performances and the four likelihoods as functions of \( \gamma \). For the likelihoods, we observe clear maxima for realistic substitution probabilities in the range of \( 0.1 \leq \gamma \leq 0.2 \) in all four cases. However, for the classification performances, we observe the four maxima for unrealistically high substitution probabilities \( \gamma \geq 0.8 \). This observation is again surprising and states that the classification performance of phylogenetic footprinting is higher for an unrealistically high substitution probability of \( \gamma = 1.0 \) than for realistic substitution probabilities in the range of \( 0.1 \leq \gamma \leq 0.4 \) for all five TFs CTCF, GABP, NRSF, SRF and STAT1.

In order to test if this result could be an artifact of the choice of the negative dataset, we study the classification performance when negatives are taken from the positives of the other datasets as done by \textit{Arnold et al.} (2012). We obtain the same surprising results that the classification performance is higher for a substitution probability of \( \gamma = 1.0 \) than for realistic substitution probabilities for all five TFs (Supplementary Figs S5, S9, S13, S17 and S21).

Next, we scrutinize the motifs obtained by PFMs with a substitution probability of \( \gamma = 1.0 \). For synthetic data, we find that the motifs obtained by PFMs with \( \gamma = 1.0 \) are highly similar to the motifs used for data generation (Supplementary Fig. S1). For real data, we find that the motifs obtained by PFMs with \( \gamma = 1.0 \) are highly similar to the motifs obtained by PFMs with realistic substitution probabilities in the range of \( 0.1 \leq \gamma \leq 0.4 \) (Supplementary Figs S2, S6, S10, S14 and S22). These findings suggest that the
Fig. 3. Classification performance and likelihood for different substitution probabilities. We plot the classification performance (decreasing) and likelihood (increasing) on data of the four TFs GABP, NRSF, SRF and STAT1 for substitution probabilities γ ∈ {0.05, 0.1, …, 1.0}. (a) GABP. The maximum likelihood is achieved for γ = 0.2. The best classification performance is achieved for γ = 1.0. (b) NRSF. Maximum likelihood is achieved for γ = 0.15. The best classification performance is achieved for γ = 0.8. (e) STAT1. The maximum likelihood is achieved for γ = 0.15. The best classification performance is achieved for γ = 1.0. (d) SRF. The maximum likelihood is achieved for γ = 0.15. The best classification performance is achieved for γ = 1.0. For each of the four TFs, we find qualitatively similar curves when quantifying the classification performance by the area under the PR curve (see Supplementary Figs S8, S12, S16 and S20).

motifs obtained by PFMs with an unrealistically high substitution probability of γ = 1.0 might be less biased than naïvely expected.

3.4 Classification performance using realistic phylogenetic trees

Fourth, we study if the phenomenon that the maximum classification performance is achieved for unrealistically high values of γ, which we observed for PFMs with a star topology, also occurs when using realistic phylogenetic trees. This study is motivated by observations that PFMs with phylogenetic trees with realistic tree topologies have the potential to yield higher classification performances than PFMs with phylogenetic trees with unrealistic star topologies (Newberg et al., 2007; Palumbo and Newberg, 2010).

Hence, we study the classification performances of PFMs on synthetic data with different tree topologies and different substitution probabilities, and we find in all cases the highest classification performances near the substitution probabilities used for data generation (Supplementary Material section 4.2 and Supplementary Fig. S25). In addition to generating synthetic data by the F81 substitution model (Felsenstein, 1981), we also generate them by the more realistic HKY substitution model Hasegawa et al. (1985) in combination with different tree topologies and different substitution probabilities, and we find again the highest classification performances near the substitution probabilities used for data generation (Supplementary Material sections 4.4 and 4.5 and Supplementary Figs S27 and S28).

Next, we study the classification performance of the PFM on real data using a phylogenetic tree and substitution probabilities from the literature (Arnold et al., 2012). We denote the PFM with a phylogenetic tree and substitution probabilities from the literature by $M_{\text{PFM}}^{\text{lit}}$, the PFM with a phylogenetic tree with a star topology and substitution probabilities according to the maximum-likelihood estimates of Figures 1b and 3a–d by $M_{\text{PFM}}^{\text{star}}$, and the PFM with a phylogenetic tree with a star topology and substitution probabilities of Figures 1b and 3a–d by $M_{\text{PFM}}^{\text{ML}}$.

Fig. 4. Classification performance of three PFMs on real data of five TFs. The PFM $M_{\text{PFM}}^{\text{star}}$ (right) outperforms the PFMs $M_{\text{PFM}}^{\text{lit}}$ (left) and $M_{\text{PFM}}^{\text{ML}}$ (middle), which implies that assuming phylogenetic independence generally improves motif prediction. The PFM $M_{\text{PFM}}^{\text{star}}$ typically achieves a higher classification performance than the PFM $M_{\text{PFM}}^{\text{ML}}$ (see Supplementary Table S3 for significances). For each of the five TFs, we find qualitatively similar results by the area under PR curve (see Supplementary Fig. S23) with similar significances shown in Supplementary Table S4. Supplementary Figures S23 also shows a comparison of $M_{\text{PFM}}^{\text{star}}$, $M_{\text{PFM}}^{\text{lit}}$ and $M_{\text{PFM}}^{\text{star}}$ with two additional PFMs (Color version of this figure is available at Bioinformatics online.)

Possible explanations for this unexpected observation might be unrealistic model assumptions—namely a phylogenetic tree with a star topology and substitution probabilities of γ = 1.0—might yield higher classification performances than the same PFMs with more realistic phylogenetic trees and more realistic substitution probabilities.

4 Discussion

Possible explanations for this unexpected observation might be unrealistic model assumptions of the substitution model, heterogeneous substitution probabilities at different TFBS positions and in different DNA regions, heterotachious substitution probabilities at different times of evolution, or the construction of incorrect or at least partially erroneous MSAs.

Violations of model assumptions sometimes lead to a poor classification performance or to a strange dependence of the classification performance on one or several model parameters. Such a situation might occur in phylogenetic footprinting, where PFMs typically assume the same phylogenetic tree and the same substitution probabilities for all positions of all TFBSs, for all TFBSs and all of their flanking regions, and for all chromosomal regions and all MSAs despite the fact that all of these assumptions are almost certainly violated (Conrad et al., 2013; Lercher and Hurst, 2002; Moses et al., 2003; Schuster-Böckler and Lehner, 2012; Tian et al., 2008; Weber et al., 2007; Wolfe et al., 1989).

Heterogeneous substitution probabilities among different DNA regions are omnipresent and typically taken into account when modeling the evolution of proteins or protein-coding genes. However, this heterogeneity is typically neglected in PFMs, where this
assumption would lead to potential over-fitting (Hawkins, 2004) due to the facts that the positions of TFBSs are unknown in phylogenetic footprinting and that TFBSs are much shorter than protein-coding genes.

Heterotachy substitution probabilities, i.e., substitution probabilities that vary with time, are another feature that is typically neglected in PFMs despite being omnipresent in both functional TFBSs as well as their flanking regions. Neglecting heterotachy might lead to the estimation of severely biased substitution probabilities, to incorrect motif predictions, and thus to a poor classification performance (Kolaczkowski and Thornton, 2004).

Incorrect or at least partially erroneous MSAs are another problem that might lead to the violation of model assumptions (Kim and Ma, 2011; Löytynoja et al., 2012). In particular, insertions and deletions as well as heterogeneity in sequence composition such as a varying GC-content (Hardison and Taylor, 2012) might cause MSA algorithms to become potentially imprecise and might thus affect all downstream analyses (Löytynoja and Goldman, 2008).

Maximum-likelihood estimators can be proven to achieve the highest classification performance in the asymptotic limit of infinitely large datasets and under the prerequisite that the models used for classification are exactly those used for data generation. However, both prerequisites are typically not fulfilled in practice, so it often happens that the highest classification performance is not achieved by those parameters that maximize the likelihood.

This situation apparently occurs for phylogenetic footprinting in a surprisingly pronounced manner, which seems to indicate that the likelihoods of currently used PFMs are less affected by violated model assumptions than their classification performances. On an intuitive level, PFMs with realistic phylogenetic trees and realistic substitution probabilities seem to be more strongly affected by heterogeneity, heterotachy and errors in MSAs than PFMs with unrealistically high substitution probabilities, so using such unrealistically high substitution probabilities might be a temporarily useful choice until more sophisticated PFMs capable of coping with heterogeneity, heterotachy and errors in MSAs are being developed.

5 Conclusions

We have studied the influence of choosing different phylogenetic trees and different substitution probabilities on the likelihood and the classification performance of PFMs. We have performed these studies on synthetic and real data obtained from ChIP-Seq experiments performed in human and MSAs of ChIP-Seq positive regions with upstream regions of orthologous genes in monkey, cow, dog and horse.

We find that the likelihood depends on the substitution probability in a qualitatively similar manner for synthetic and real data, where it reaches a maximum for realistic substitution probabilities in the range of $0.1 \leq \gamma \leq 0.2$. In contrast, we find that the classification performance depends on the substitution probability in a qualitatively different manner for synthetic and real data.

For synthetic data, the classification performance reaches a maximum at the values of the substitution probability used for data generation, which coincide with those values that maximize the likelihood. For real data, however, it increases with the substitution probability and stops increasing only at unrealistically high values of the substitution probability in the range of $0.8 \leq \gamma \leq 1$, which are very different from those values that maximize the likelihood.

We find in all of the studied datasets that PFMs using unrealistic substitution probabilities of $\gamma = 1.0$ yield higher classification performances than PFMs using realistic substitution probabilities.

One possible explanation for this strange behavior of the classification performance on the substitution probability is the presence of heterogeneous and heterotachious substitution probabilities, which are neglected by currently used PFMs, and the sensitive dependence of PFMs on the reconstructed MSAs that might be partially incorrect.

Apparently, PFMs using unrealistic substitution probabilities of $\gamma = 1.0$ are more robust to these and possibly other violations of the model assumptions than PFMs based on realistic substitution probabilities, and this robustness might lead to less biased parameter estimates and thus more accurate phylogenetic footprints.

This observation leads to the strange practical recommendation of using PFMs using unrealistic substitution probabilities of $\gamma = 1.0$ instead of using PFMs using realistic substitution probabilities until there are more sophisticated models for the evolution of TFBSs and their flanking regions that take into account heterogeneity and heterotachy as well as partially erroneous alignments in a position-specific manner.

Acknowledgements

We thank Karin Breunig, Ralf Eggeling, Jan Grau, and Peter Stadler for valuable discussions.

Funding

We thank DFG [grant no. GR3526/1] for financial support.

Conflict of Interest: none declared.

References
