Bioinformatics, 33(11), 2017, 1639–1646
doi: 10.1093/bioinformatics/btx033
Advance Access Publication Date: 27 January 2017
Original Paper

OXFORD

Phylogenetics

Unrealistic phylogenetic trees may improve phylogenetic footprinting

Martin Nettling[1,*], Hendrik Treutler[2], Jesus Cerquides[3] and Ivo Grosse[1,4]

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

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on March 2, 2016; revised on December 2, 2016; editorial decision on January 18, 2017; accepted on January 19, 2017

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 promotor 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

biology, medicine and biodiversity research (Nowrousian, 2010; Villar *et al.*, 2014).

The last decade has witnessed a spectacular development of sequencing technologies unleashing new potentials in identifying TFBSs (Kulakovskiy *et al.*, 2010; Furey, 2012; Lasken and McLean, 2014; van Dijk *et al.*, 2016). Due to the increasing number of available 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; Schaefke *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_n.

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 conform a tree. Thus, for each species k but the root, $pa(k)$ denotes the ancestor of species k in the tree. The root species is noted r.

Our probabilistic model contains a random variable $S_n^{u,k}$ 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 $\mathcal{A} = \{A, C, G, T\}$. We note $\mathrm{pa}(S_n^{u,k})$ the uth nucleotide in the nth alignment of species $\mathrm{pa}(k)$ (the ancestor of k). By definition, the root has no ancestor and hence $\mathrm{pa}(S_n^{u,r}) = \varnothing$. We also refer to nucleotide $S_n^{u,k}$ as $A_n^{u,k}$ when species k is observed, and as $Y_n^{u,k}$ when species k is hidden. Furthermore we note by $Y_n^{u,\cdot}$ (respectively $S_n^{u,\cdot}$) the set containing each random variable $Y_n^{u,k}$ (respectively $S_n^{u,k}$), with $O+1 \leq k \leq O+H$ and Y_n the set containing every random variable in $Y_n^{u,\cdot}$ 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^0 indicating that alignment A_n does not contain a motif and M_n^1 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

$$p(A_n|\theta) = p(A_n|M_n^0, \theta) \times p(M_n^0|\theta) + p(A_n|M_n^1, \theta) \times p(M_n^1|\theta)$$

with variable M_n taking a Bernoulli distribution and θ 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 $p(A_n|M_n^0,\theta)$ and for motif-bearing alignments $p(A_n|M_n^1,\theta)$. For reasons of clarity we omit θ 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

$$p(A_n|M_n^0) = \sum\nolimits_{Y_n} p(A_n|Y_n, M_n^0). \tag{1}$$

We assume that each single nucleotide alignment is independent of any other nucleotide alignment given θ and M_n^0. Furthermore, we assume that in each nucleotide alignment, the species satisfy the conditional independencies encoded by the phylogenetic tree. Thus,

$$p(A_n|M_n^0) = \prod_{u=1}^{L_n} \sum\nolimits_{Y_n^{u,\cdot}} p(S_n^{u,\cdot}|M_n^0) \tag{2}$$

$$= \prod_{u=1}^{L_n} \sum\nolimits_{Y_n^{u,\cdot}} \prod_{k=1}^{O+H} p(S_n^{u,k}|\mathrm{pa}(S_n^{u,k}), M_n^0) \tag{3}$$

where

$$p(S_n^{u,k} = a|\mathrm{pa}(S_n^{u,k}) = b, M_n^0) = \begin{cases} \pi_0^a & \text{if } k = r \\ \gamma_k \times \pi_0^a + (1-\gamma_k)\delta_{a=b} & \text{if } k \neq r \end{cases}$$

according to the F81 model, where the base distribution of each position of the background sequence is denoted by π_0, the probability of a nucleotide a in the background sequence is denoted by π_0^a, and the substitution probability from the ancestor species to species k is denoted by γ_k. For more realistic phylogenetic models γ_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

$$p(A_n|M_n^1) = \sum\nolimits_{\ell_n=1}^{L_n-W+1} \sum\nolimits_{Y_n} p(A_n, Y_n, \ell_n|M_n^1). \tag{4}$$

where W is the length of the TFBS and ℓ_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

$$p(A_n|M_n^1) = \sum_{\ell_n=1}^{L_n-W+1} p(\ell_n|M_n^1) \prod_{u=1}^{L_n} \sum_{Y_n^{u,\cdot}} p(S_n^{u,\cdot}|\ell_n, M_n^1) \tag{5}$$

with $p(S_n^{u,\cdot}|\ell_n, M_n^1) = \prod_{k=1}^{\mathrm{O+H}} p(S_n^{u,k}|\mathrm{pa}(S_n^{u,k}), \ell_n, M_n^1)$ and

$$p(S_n^{u,k}|pa(S_n^{u,k}), \ell_n, M_n^1) = \begin{cases} \pi_0^a & \text{if } k = r \text{ and } u < \ell_n \text{ or } u \geq \ell_n + W \\ \pi_{u-\ell_n+1}^a & \text{if } k = r \text{ and } \ell_n \leq u < \ell_n + W \\ \gamma_k \times \pi_0^a + (1-\gamma_k)\delta_{a=b} & \text{if } k \neq r \text{ and } u < \ell_n \text{ or } u \geq \ell_n + W \\ \gamma_k \times \pi_{u-\ell_n+1}^a + (1-\gamma_k)\delta_{a=b} & \text{if } k \neq r \text{ and } \ell_n \leq u < \ell_n + W \end{cases}$$

As for the non-motif-bearing alignment, the base distribution of each position of the background sequence is denoted by π_0 and the probability of a nucleotide a in the background sequence is denoted by π_0^a. The base distributions of a motif sequence of length W are denoted by π_w with $w \in [1,\ldots,W]$ and the probability of a nucleotide a at position w in a motif sequence is denoted by π_w^a. The substitution probability from the ancestor species to species k is denoted by γ_k.

Finally we assume motifs to be uniformly distributed, thus having that $p(\ell_n|M_n^1) = \frac{1}{L_n-W+1}$, 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 Jothi *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 π_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 π_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 sequence, we sample the nucleotide at position u from the uniform distribution π_0. For each position $u \in [\ell, \ell + W - 1]$ of the TFBS, we sample the nucleotide at position u from the distribution $\pi_{u-\ell_n+1}$. The distribution π_w with $w \in \{1,\ldots,15\}$ 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 π_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 $\pi_{u-\ell_n+1}$ 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 γ is inversely proportional to the phylogenetic relatedness. Small γ encode close phylogenetic relations and large γ 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/.

Algorithm 1. Motif discovery algorithm for the proposed PFM. Upon random initialization of the model parameters we iteratively estimate sequence weights and model parameters in multiple algorithm restarts, where R denotes the number of restarts of the whole algorithm, and T denotes the number of iterations. The result is the set of model parameters together with maximum likelihood.

Data: Set of alignments $A = \{A_1, \ldots, A_N\}$
Flanking model: Maximize $p(A|\theta^1)$ for the model parameters $\pi_0 \subset \theta^1$
for r = 1 ...R **do**
  Initialize $\pi_w \subset \theta^1$ randomly for $w \in \{1, \ldots, W\}$
  for t = 1 ...T **do**
    E-step: Estimate $p(A_n|\ell_n, M_n^1, \theta^t)$ for each position ℓ_n in each alignment A_n given the model parameters θ^t
    M-step: Maximize the expected value of the complete-data log-likelihood with respect to model parameters π_w and denote the resulting argmax by θ^{t+1}.
  end for
  Keep θ^{T+1} denoted θ_r
end for
Result: $\theta \in \{\theta_1, \ldots \theta_R\}$ with maximum likelihood

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 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 γ 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 γ 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 γ of the PFMs is in the range of $0.1 \leq \gamma \leq 0.4$ according to Gertz *et al.* (2006).

Figure 1a shows the likelihood as a function of the substitution probability γ 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 γ 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 Gertz *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 γ from 0.05 to 1.0 with increments of 0.05 and compute the classification performance as a function of γ 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 γ, i.e. we expect that the classification performance is highest for γ 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.

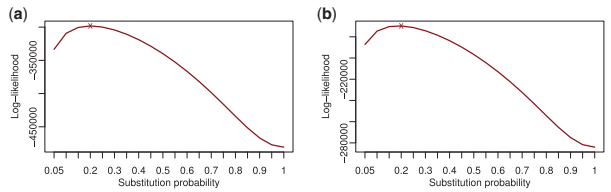

Fig. 1. Likelihood for different substitution probabilities. We plot the likelihood on synthetic data and CTCF data for a PFM using a star topology with all substitution probabilities set to $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$. **(a)** Synthetic data. Maximum likelihood is achieved for $\gamma = 0.2$, the substitution probability used for data generation. **(b)** CTCF data. Maximum likelihood is achieved for $\gamma = 0.2$, lying in the range of $0.1 \leq \gamma \leq 0.4$ suggested by the literature

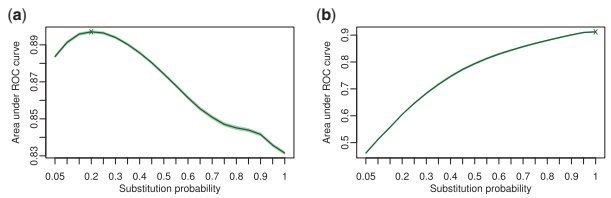

Fig. 2. Classification performance for different substitution probabilities. We plot the classification performance on synthetic data and CTCF data for a PFM using a star topology with all substitution probabilities set to $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$. **(a)** Synthetic data. Highest classification performance is achieved for $\gamma = 0.25$, which is close to $\gamma = 0.2$, the substitution probability used for data generation. **(b)** CTCF data. Highest classification performance is achieved for $\gamma = 1.0$, which is unrealistic and different from the expected result. We obtain similar results when quantifying the classification performance by the area under the PR curve (Supplementary Fig. S4)

Figure 2a shows the classification performance as a function of γ 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 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 γ 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 Gertz *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 unrealistic 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 γ 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 γ. 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 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

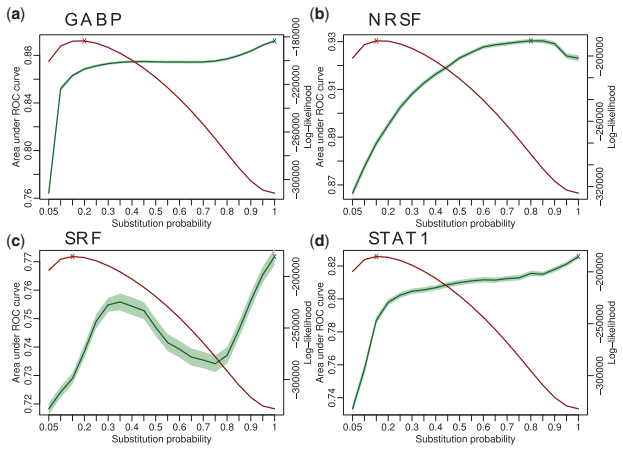

Fig. 3. Classification performance and likelihood for different substitution probabilities. We plot the classification performance (decreasing) and likelihood (increasing) on data of the four TFs GABP, NRSFm, SRF and STAT1 for substitution probabilities $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$. **(a)** GABP. The maximum likelihood is achieved for $\gamma = 0.2$. The best classification performance is achieved for $\gamma = 1.0$. **(b)** NRSF. Maximum likelihood is achieved for $\gamma = 0.15$. The best classification performance is achieved for $\gamma = 0.8$. **(c)** STAT1. The maximum likelihood is achieved for $\gamma = 0.15$. The best classification performance is achieved for $\gamma = 1.0$. **(d)** SRF. The maximum likelihood is achieved for $\gamma = 0.15$. The best classification performance is achieved for $\gamma = 1.0$. 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 $\gamma = 1.0$ might be less biased than naively 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 $\mathcal{M}_{\text{lit}}^{\text{tree}}$, 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 $\mathcal{M}_{\text{ML}}^{\text{star}}$, and the PFM with a

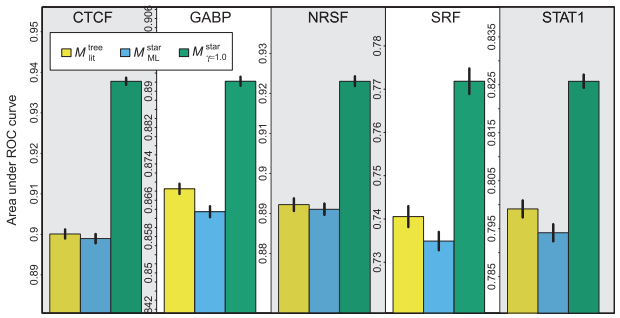

Fig. 4. Classification performance of three PFMs on real data of five TFs. The PFM $\mathcal{M}_{\gamma=1.0}^{\text{star}}$ (right) outperforms the PFMs $\mathcal{M}_{\text{lit}}^{\text{tree}}$ (left) and $\mathcal{M}_{\text{ML}}^{\text{star}}$ (middle), which implies that assuming phylogenetic independence generally improves motif prediction. The PFM $\mathcal{M}_{\text{lit}}^{\text{tree}}$ typically achieves a higher classification performance than the PFM $\mathcal{M}_{\text{ML}}^{\text{star}}$ (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 $\mathcal{M}_{\gamma=1.0}^{\text{star}}$, $\mathcal{M}_{\text{ML}}^{\text{star}}$ and $\mathcal{M}_{\text{lit}}^{\text{tree}}$ with two additional PFMs (Color version of this figure is available at *Bioinformatics* online.)

phylogenetic tree with a star topology and substitution probabilities of $\gamma = 1.0$ by $\mathcal{M}_{\gamma=1.0}^{\text{star}}$.

Figure 4 shows the classification performances of $\mathcal{M}_{\text{lit}}^{\text{tree}}$, $\mathcal{M}_{\text{ML}}^{\text{star}}$ and $\mathcal{M}_{\gamma=1.0}^{\text{star}}$ for each of the five TFs CTCF, GABP, NRSF, SRF and STAT1. Interestingly, we find that $\mathcal{M}_{\gamma=1.0}^{\text{star}}$ yields a significantly higher classification performance than the other two PFMs. In addition, we investigate the classification performances of PFMs with a star topology and a tree topology from the literature with branch lengths estimated from the data, and we find also in this case that $\mathcal{M}_{\gamma=1.0}^{\text{star}}$ yields a significantly higher classification performance than the other two PFMs (Supplementary Material section 3 and Supplementary Fig. S23).

These findings state that, in case of real data, choosing unrealistic model assumptions—namely a phylogenetic tree with a star topology and substitution probabilities of $\gamma = 1.0$—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.*, 2011; 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.

Heterotachious 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 by 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

Anisimova,M. *et al.* (2013) State-of the art methodologies dictate new standards for phylogenetic analysis. *BMC Evolution. Biol.*, **13**, 161.

Arnold,P. *et al.* (2012) Motevo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of dna sequences. *Bioinformatics*, **28**, 487–494.

Blanchette,M. and Tompa,M. (2003) Footprinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res.*, **31**, 3840–3842.

Clark,A.G. *et al.* (2007) Evolution of genes and genomes on the drosophila phylogeny. *Nature*, **450**, 203–218.

Conrad,D.F. *et al.* (2011) Variation in genome-wide mutation rates within and between human families. *Nature*, **43**, 712–714.

de Fougerolles,A. *et al.* (2007) Interfering with disease: a progress report on sirna-based therapeutics. *Nat. Rev. Drug Discov.*, **6**, 443–453.

Doniger,S.W. and Fay,J.C. (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput. Biol.*, **3**, e99.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Furey,T.S. (2012) ChIPseq and beyond: new and improved methodologies to detect and characterize proteinDNA interactions. *Nat. Rev. Genet.*, **13**, 840–852.

Gertz,J. *et al.* (2006) Phylogeny based discovery of regulatory elements. *BMC Bioinformatics*, **7**, 266.

Grau,J. *et al.* (2012) Jstacs: a java framework for statistical analysis and classification of biological sequences. *J. Mach. Learn. Res.*, **13**, 1967–1971.

Hardison,R.C. and Taylor,J. (2012) Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.*, **13**, 469–483.

Hasegawa,M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J. Mol. Evol.*, **22**, 160–174.

Hawkins,D.M. (2004) The problem of overfitting. *J. Chem. Inform. Comput. Sci.*, **44**, 1–12.

Hawkins,J. *et al*. (2009) Assessing phylogenetic motif models for predicting transcription factor binding sites. *Bioinformatics*, **25**, i339–i347.

Hobert,O. (2008) Gene regulation by transcription factors and micrornas. *Science*, **319**, 1785–1786.

Jothi,R. *et al*. (2008) Genome-wide identification of in vivo protein-dna binding sites from chip-seq data. *Nucl. Acids Res*., **36**, 5221–5231.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. *Mammal. Protein Metab*., **3**, 132.

Katoh,K. and Standley,D.M. (2013) Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol*., **30**, 772–780.

Kc,D.B. and Livesay,D.R. (2011) Topology improves phylogenetic motif functional site predictions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. **8**, 226–233.

Kim,J. and Ma,J. (2011) Psar: measuring multiple sequence alignment reliability by probabilistic sampling. *Nucleic Acids Res*., **39**, 6359–6368.

Kolaczkowski,B. and Thornton,J.W. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, **431**, 980–984.

Kulakovskiy,I.V. *et al*. (2010) Deep and wide digging for binding motifs in chip-seq data. *Bioinformatics*, **26**, 2622–2623.

Lasken,R.S. and McLean,J.S. (2014) Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat. Rev. Genet*., **15**, 577–584.

Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.

Lercher,M.J. and Hurst,L.D. (2002) Human snp variability and mutation rate are higher in regions of high recombination. *Trends Genet*., **18**, 337–340.

Löytynoja,A. and Goldman,N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632–1635.

Löytynoja,A. and Goldman,N. (2010) webprank: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*, **11**, 579.

Löytynoja,A. *et al*. (2012) Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics*, **28**, 1684–1691.

Luco,R.F. *et al*. (2010) Regulation of alternative splicing by histone modifications. *Science*, **327**, 996–1000.

Moses,A.M. *et al*. (2004a) Monkey: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*., **5**, R98.

Moses,A.M. *et al*. (2003) Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol. Biol*., **3**, 19.

Moses,A.M. *et al*. (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pacific Symposium on Biocomputing*. Hawaii, United States, pp. 324–335.

Nettling,M. *et al*. (2017) Combining phylogenetic footprinting with motif models incorporating intra-motif dependencies. *BMC Bioinformatics* (In press).

Newberg,L.A. *et al*. (2007) A phylogenetic gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics*, **23**, 1718–1727.

Notredame,C. *et al*. (2000) T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol*., **302**, 205–217.

Nowrousian,M. (2010) Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryot. Cell*, **9**, 1300–1310.

Palumbo,M.J. and Newberg,L.A. (2010) Phyloscan: locating transcription-regulating binding sites in mixed aligned and unaligned sequence data. *Nucleic Acids Res*., **38**, W268–W274.

Pollard,K.S. *et al*. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*., **20**, 110–121.

Sainsbury,S. *et al*. (2015) Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol*., **16**, 129–143.

Schaefke,B. *et al*. (2015) Gains and losses of transcription factor binding sites in saccharomyces cerevisiae and saccharomyces paradoxus. *Genome Biol. Evol*., **7**, 2245–2257.

Schoenberg,D.R. and Maquat,L.E. (2012) Regulation of cytoplasmic mRNA decay. *Nat. Rev.. Genet*., **13**, 246–259.

Schuster-Böckler,B. and Lehner,B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, **488**, 504–507.

Siddharthan,R. (2008) Phylogibbs-mp: module prediction and discriminative motif-finding by gibbs sampling. *PLoS Comput. Biol*., **4**, e1000156.

Siddharthan,R. *et al*. (2005) PhyloGibbs: a gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol*., **1**, e67.

Sinha,S. *et al*. (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, **5**, 170.

Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat. Rev. Genet*., **14**, 204–220.

Sultan,M. *et al*. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.

Tam,O.H. *et al*. (2008) Pseudogene-derived small interfering rnas regulate gene expression in mouse oocytes. *Nature*, **453**, 534–538.

Tessarz,P. and Kouzarides,T. (2014) Histone core modifications regulating nucleosome structure and dynamics. *Nat. Rev. Mol. Cell Biol*., **15**, 703–708.

Tian,D. *et al*. (2008) Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature*, **455**, 105–108.

Tuğrul,M. *et al*. (2015) Dynamics of transcription factor binding site evolution. *PLoS Genet*., **11**, e1005639.

Valouev,A. *et al*. (2008) Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat. Methods*, **5**, 829–834.

van Dijk,E.L. *et al*. (2016) Ten years of next-generation sequencing technology. *Trends Genet*., **30**, 418–426.

Villar,D. *et al*. (2014) Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat. Rev. Genet*., **15**, 221–233.

Voss,T.C. and Hager,G.L. (2014) Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat. Rev. Genet*., **15**, 69–81.

Weber,M. *et al*. (2007) Distribution, silencing potential and evolutionary impact of promoter dna methylation in the human genome. *Nat. Genet*., **39**, 457–466.

Wolfe,K.H. *et al*. (1989) Mutation rates differ among regions of the mammalian genome. *Nature*, 283–285. pages