An Empirical Prior Improves Accuracy for Bayesian Estimation of Transcription Factor Binding Site Frequencies within Gene Promoters

Stephen A. Ramsey[1,2]

[1]Department of Biomedical Sciences, Oregon State University, Corvallis, OR, USA. [2]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.

Supplementary Issue: Current Developments in Domestic Animal Bioinformatics

ABSTRACT: A Bayesian method for sampling from the distribution of matches to a precompiled transcription factor binding site (TFBS) sequence pattern (conditioned on an observed nucleotide sequence and the sequence pattern) is described. The method takes a position frequency matrix as input for a set of representative binding sites for a transcription factor and two sets of noncoding, 5′ regulatory sequences for gene sets that are to be compared. An empirical prior on the frequency A (per base pair of gene-vicinal, noncoding DNA) of TFBSs is developed using data from the ENCODE project and incorporated into the method. In addition, a probabilistic model for binding site occurrences conditioned on λ is developed analytically, taking into account the finite-width effects of binding sites. The count of TFBS β (conditioned on the observed sequence) is sampled using Metropolis–Hastings with an information entropy-based move generator. The derivation of the method is presented in a step-by-step fashion, starting from specific conditional independence assumptions. Empirical results show that the newly proposed prior on β improves accuracy for estimating the number of TFBS within a set of promoter sequences.

KEYWORDS: transcription factor, binding site, Bayesian statistics, enrichment analysis, gene regulation

SUPPLEMENT: Current Developments in Domestic Animal Bioinformatics

CITATION: Ramsey. An Empirical Prior Improves Accuracy for Bayesian Estimation of Transcription Factor Binding Site Frequencies Within Gene Promoters. *Bioinformatics and Biology Insights* 2015:9(S4) 59–69 doi: 10.4137/BBI.S29330.

TYPE: Methodology

RECEIVED: May 11, 2016. **RESUBMITTED:** September 11, 2016. **ACCEPTED FOR PUBLICATION:** September 18, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Eight peer reviewers contributed to the peer review report. Reviewers' reports totaled 1158 words, excluding any confidential comments to the academic editor.

FUNDING: This study was supported by award number K25HL098807 from the National Heart, Lung and Blood Institute, the Medical Research Foundation of Oregon (New Investigator Grant award), the National Science Foundation (award numbers 1557605-DMS and 1553728-DBI), the PhRMA Foundation (Research Starter Grant), and Oregon State University (Division of Health Sciences Interdisciplinary Research Grant award).

The author confirms that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Author discloses no potential conflicts of interest.

CORRESPONDENCE: stephen.ramsey@oregonstate.edu

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

In bioinformatics, an enduring and fundamental question is how best to use an organism's genome sequence as well as prior knowledge of the DNA sequence preferences of transcription factors (TFs) in order to determine which TFs are responsible for an observed pattern gene expression differences between sample groups, such as tissues at different stages of disease and cells cultured in the presence or absence of a chemical stimulus.[1–3] The general approach of computationally analyzing noncoding DNA sequences within 5′ (upstream) regions of differentially expressed gene sets to identify statistically overrepresented TF binding site (TFBS) sequence matches – known as TFBS enrichment analysis[4–8] – has proved useful for identifying the gene regulatory mechanisms from transcriptome data.[9–14] Databases such as MatBase,[15] TRANSFAC,[16] JASPAR,[17] UniPROBE,[18] Factorbook,[19] and FootprintDB[20] are rapidly accumulating position-nucleotide frequency matrices (PFMs) that represent the sequence preferences of individual TFs. This rapid accumulation is driven by high-throughput assays such as ChIP-seq and protein-binding microarrays and through the use of improved *in silico* structural models for predicting TF–DNA affinities. Although a ChIP-seq assay can be used to map the binding sites of a specific TF genome-wide within a specific cell type or tissue,[21] only a small percentage of known TFs have been successfully assayed using this technique. In vertebrates, there have been relatively few reports of applications of ChIP-seq outside of humans and model species such as mouse.[22,23] Thus, the approach of computationally analyzing a set of 5′ regulatory sequences to measure the enrichment of TFBS – leveraging databases of *known* TFBS sequence patterns – remains unmatched in terms of the number of TFBS sequence patterns that can be simultaneously analyzed. This discovery power is particularly important in vertebrates, for which there are ~1800 different TFs, of which hundreds can be expressed in any given cell type or tissue.[24]

Reflecting the importance of this problem, multiple computational approaches have been proposed for PFM-guided detection of enrichment of TFBS within gene-vicinal sequences.[7,25–27] For the purpose of specificity, I define *gene-vicinal* to mean within approximately 5 kbp (in either direction) of a transcription start site.[28] The TFBS enrichment analysis method of Frith et al.[7] involves the direct use of the

position-probability matrix (PPM, which is the row-normalized PFM) in order to compute a likelihood ratio of the PPM model to a nucleotide frequency-based background model, for a binding site-sized sequence window at a given position. The likelihood ratios are then averaged over all nucleotide positions within a single gene-vicinal sequence to obtain a single-gene score. For each possible subset of genes from the gene set, the product of gene-level scores is computed, and these subset-level scores are averaged.[7] In another approach, Ho Sui et al.[25] used a log-likelihood ratio approach with an empirically determined hard threshold in order to identify TFBS and then used the binomial distribution to test the enrichment of TFBS. Sinha and Tompa[29] used a multi-TF approach in which the weighted sum of occurrences of a specific TF's PPM was computed over binding site configurations for all TF PPMs to be analyzed. The prior on the expected number of binding sites is not treated probabilistically but is a fixed parameter value. Pavesi and Zambelli[27] rescaled the positional log-likelihood score in order to map the score to a compact interval and then computed the maximum of this rescaled score at all positions within a gene-vicinal sequence; this per-gene score is then averaged over all genes in the gene set. The diversity of methods for PFM-guided TFBS enrichment analysis and the significant numbers of studies (over 600 combined, for Refs. 7 and 25) that have reported using these methods underscores the importance of this problem in the field of bioinformatics.

Despite its discovery power, TFBS enrichment analysis using prior TF binding pattern information in the form of PFMs has a fundamental challenge that PFMs are highly variable in terms of their specificity for nucleotide sequences and in terms of the uncertainty of the composition of the corresponding PPMs.[30,31] Within databases of TFBS sequence patterns, the numbers of representative binding sites from which individual vertebrate TF PFMs have been compiled can vary by four orders of magnitude, from half a dozen to tens of thousands of representative oligomer sequences.[15–17] For cases of TFs with highly specific nucleotide affinity and/or very low sampling of representative binding site sequences, PFM counts of zero pose a problem in the standard PPM-based approach and necessitate the use of ad hoc pseudocounts to enable the scoring of nucleotide sequences that do not perfectly match the TFBS consensus sequence.[32,33] Furthermore, because the precision of the PPM that is associated with a PFM depends directly on the number of representative binding site sequences used to compile the PFM,[30] TFBS enrichment analysis using only the PPM (and not taking into account the uncertainty in the PPM's structure) can be a source of both type I and type II errors. Finally, in order to assess the significance of a finding that the frequency of PPM sequence matches for a TF is statistically overrepresented for 5′ upstream sequences for a gene set versus for a background set of genes, it is necessary to quantify the *magnitude* of the frequency enrichment and not just statistical significance (eg, using a *P* value). In addition, it is useful to be able to estimate the uncertainty on the magnitude of the TFBS frequency enrichment. A Bayesian approach to TFBS frequency estimation, as described below, has the potential to address the challenge of highly variable accuracy (sharpness) of known TFBS motifs. Bayesian methods have long been used for de novo motif discovery[34–37] and have also been proposed for TFBS recognition and demonstrated to have improved accuracy over traditional motif scanning.[30] In the context of PFM-guided enrichment analysis, a Bayesian approach is appealing because it could account for uncertainty in the PPM and it could provide an estimate of the TFBS frequency per base pair of noncoding DNA, while appropriately weighting high-quality and low-quality matches to the PPM. By using a Bayesian approach, an additional benefit is that an empirical prior distribution of TFBS frequencies (across many TFs) can be included in the model to improve the TFBS frequency estimation in the case of a weak (ie, degenerate) PPM.

In this article, I describe a Bayesian approach to PFM-guided TFBS enrichment analysis, which produces samples from the posterior distribution of the number of TFBS for a given PFM, within a given sequence. The method incorporates an empirical prior on the per base pair TFBS frequency that is informed by the analysis of human TFBS from the ENCODE project (as opposed to the geometric prior used in a previous study[30]). Finally, because the method is developed from an explicit joint probability model of all of the observables and model parameters, the method could be readily extended to incorporate other types of regulatory potential scores.[38,39] I show empirical results from applying the new prior to estimate the number of TFBS for a synthetic set of promoter sequences in which representative TFBS sequences are introduced. The empirical results show that the new prior improves accuracy when compared to a previously proposed prior on the per-promoter number of TFBS.

Mathematical Preliminaries and Notation

For the purpose of detecting TFs that may control a given cluster of coexpressed genes, it is simplest to consider a single TF at a time; I use "TFX" as a generic symbol for this TF. (Although this article is focused on single-TF enrichment analysis for simplicity, the pairwise TF enrichment analysis could in principle be accommodated by extensions of the general approach described herein.) Consistent with a Bayesian approach, I start by framing the problem of PFM-based TFBS enrichment analysis in terms of a set of random variables including observations, nuisance parameters, and a single parameter – the number of TFBSs within a given set of gene promoters – whose distribution (conditioned on the observations) will ultimately be sampled. To do so, I introduce a bit of mathematical notation needed to define these random variables. It is convenient to denote the set of natural DNA nucleotides by integers, $\mathbb{D} = \{1,2,3,4\}$, corresponding to A, C, G, and T (so the complementary nucleotide for nucleotide $d \in \mathbb{D}$ is $5 - d$). For simplicity, let the promoter sequences of a cluster of differentially expressed genes be concatenated and represented as a

sequence $\boldsymbol{s} \in \mathbb{D}^L$, where L is the total sequence length in base pair. Let the noncoding DNA sequence within the promoters of a set of randomly selected genes that are expressed (but not necessarily differentially expressed) within the same cell type or tissue, be represented by $\boldsymbol{s''} \in \mathbb{D}^{L''}$, where L'' is the length in base pair. Finally, let $\boldsymbol{s'} \in \mathbb{D}^{L'}$ be a large DNA sequence comprising noncoding, gene-vicinal sequences selected at random and in which any known TFBS (as mapped by high-throughput ChIP-seq studies) have been excluded (here again, L' is the total length, in base pair). The *background model* sequence $\boldsymbol{s'}$ will be used to obtain a model for nucleotide frequencies in noncoding, nonbinding site DNA. The *regulatory region* sequences $\boldsymbol{s}$ and $\boldsymbol{s''}$ will be analyzed for a relative enrichment of TFBSs for a given TF, as described below.

The TFX is assumed to have a set of representative binding site sequences (numbering c; depending on the type of assay used to compile the representative binding site sequences, c could range from 6 to 100,000, as shown in Fig. 1) obtained from the literature and/or from high-throughput protein–DNA binding measurements. The representative binding site sequences are assumed to have been multiply aligned; I denote by w the length (in base pair) of the core region of overlapping representative binding sites in the multiple alignment. The counts of nucleotides of each type at each position within a PFM will be denoted by a matrix $\boldsymbol{c} \in \mathbb{C}^{w\times 4}$, where $\mathbb{C} = \{0, 1, 2, \ldots, c\}$. I note that $\boldsymbol{c}$ and w are specific to the transcription factor TFX, and this dependence could be denoted by $\boldsymbol{c}$(TFX) and w(TFX); however, for the simplicity of notation, I use the more compact $\boldsymbol{c}$ and w. The height of the TF matrix, w, can vary significantly from TF to TF; across all 4528 matrices in the TRANSFAC 2015 Professional database, w varies from 5 to 30, with a median of 12. The index set for sequence positions within a binding site for TFX will be denoted by $\mathbb{G} = (1, \ldots, w)$. The factor TFX is assumed to have an overall frequency per base pair, represented by λ, of binding sites in a given sequence of noncoding, gene-vicinal DNA; I represent our uncertainty about λ by treating it as a random variable Λ on $(0, \lambda_{max}]$, where a fixed value $\lambda_{max} \in (0,1)$ is chosen as an upper limit. An absolute physical upper limit on λ_{max} would $1/w$, given the requirement that binding sites for TFX not overlap. However, TFBSs in mammals are in general sparsely distributed, even within gene-vicinal regions[42]; thus, the value $\lambda_{max} = 0.001$ bp^{-1} is used here (and as seen in the "**Modeling $P(\boldsymbol{\beta}|\lambda)p(\lambda)$**" section, of all the TFs in the ENCODE human ChIP-seq dataset,[43] none has a binding site frequency per base pair in gene-vicinal sequence that exceeds 0.001 bp^{-1}). Because the actual locations of the TFX binding sites within a given noncoding, gene-vicinal sequence of length L are not known, I denote the binding site locations by a $\{-1,0,1\}^L$-valued random variable B. Specifically, for the outcome $\boldsymbol{B} = \boldsymbol{\beta} \in \{-1,0,1\}^L$, at each location $l \in \{1, \ldots, L\}$,

$$\beta_l = \begin{cases} 1 & \text{if there is a plus-orientation TFBS whose } 5' \text{ end is at location } l \\ 0 & \text{if there is no TFBS whose } 5' \text{ end is at location } l \\ -1 & \text{if there is a minus-orientation TFBS whose } 5' \text{ end is at location } l, \end{cases} \quad (1)$$

where in the above, +(plus)-orientation means that the PPM pattern derived from c matches the sequence on the forward strand, and –(minus)-orientation means that the PPM pattern matches the sequence on the reverse strand. The use of a discrete parameter to represent the presence/absence of a TFBS at a given nucleotide position (rather than a continuous parameter) makes the space of TFBS configurations more tractable to explore,[30,31] without sacrificing the ability to differentially weight a poor-affinity binding site from a high-affinity binding site in the analysis. The total number of binding sites in a given binding site configuration $\boldsymbol{\beta}$ is obtained by the L^1 norm, $\beta \equiv ||\boldsymbol{\beta}||_1$. For simplicity of notation, only binding site configurations β for which the entire binding site is contained within the range of sequence positions $\mathbb{L}_r$ and for which no two binding sites are overlapping by any number of nucleotides (even if the two binding sites have opposite orientations) will be allowed. Thus, the range of $\boldsymbol{B}$ is not the entirety of $\{-1,0,1\}^L$, but a subset $\mathbb{B} \subset \{-1,0,1\}^L$ defined by the above constraints. In the Bayesian approach to TFBS enrichment analysis that described below, $B = ||\boldsymbol{B}||_1$ is the integer-valued random variable whose distribution (conditioned on the observed regulatory region sequences and on the PFM for representative binding sites) is sought.

Importantly, the probability distribution on the number of binding sites will depend on the length of the DNA sequence

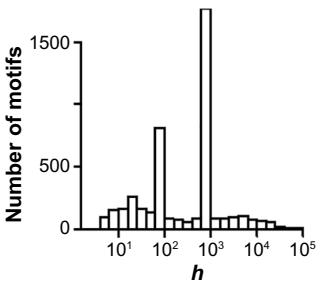

Figure 1. The distribution of values of c for a collection of 4,528 vertebrate transcription factors from the TRANSFAC Professional 2013.4 and JASPAR 5.0 databases. The sharp peak in the distribution at $c = 100$ is due to the inclusion of motif matrix information for which the original sequence alignments are not available (such as from high-throughput *in vitro* protein–DNA binding screens,[40] for which a default value of $c = 100$ was selected based on consistency with the number of significant digits for the values reported in the motif matrices). The sharp peak at $c = 998$ is due to the 2,076 structure-derived motifs that were originally obtained using the 3DTF tool[41] and were then incorporated into the TRANSFAC database). The long tail of c values above 10^3 represents motif matrices compiled from high-throughput TF location assays such as ChIP-array and ChIP-seq.

being analyzed (longer combined regulatory sequences will, in general, contain more binding sites of a given type), and thus, the probability distribution for $||\boldsymbol{B}||_1$ (the number of binding sites) conditioned on the DNA sequence $\boldsymbol{s}$ cannot be directly compared to the probability distribution for $\boldsymbol{B}|\boldsymbol{s}''$ unless $L = L''$. Thus, in practice, one would compare samples of $B/L_r|\boldsymbol{s}$ with samples of $B/L''|\boldsymbol{s}''$, with Λ treated as a nuisance variable. A key benefit of a Bayesian approach is that it will not require a specific value for the parameter λ; all possible values (consistent with the imposed constraint $\lambda_{\max}$) are considered.

A Bayesian approach to analyzing whether binding sites for TFX are enriched within sequences $\boldsymbol{s}'$ versus $\boldsymbol{s}''$ can now be succinctly described as comparing samples from the distribution of

$$B / L_r \mid c, s, s' \tag{2}$$

with samples from the distribution of

$$B / L'' \mid c, s'', s', \tag{3}$$

with Λ marginalized, under an explicit probability model. Thus, the technical problem to be solved here is how to accurately sample from the conditional distribution

$$B \mid c, r, s', \tag{4}$$

where $\boldsymbol{r}$ is an arbitrary observed set of (concatenated) promoter sequences (and in practice, one set of samples would be generated for the case $\boldsymbol{r} = \boldsymbol{s}$ and one set of samples would be generated for the case $\boldsymbol{r} = \boldsymbol{s}''$). I denote the length of the sequence $\boldsymbol{r}$ by L_r (which will have the value L or L'' depending on whether we are modeling the case $\boldsymbol{r} = \boldsymbol{s}$ or the case $\boldsymbol{r} = \boldsymbol{s}''$), and the sequence of unique positions within the combined gene-vicinal DNA regions by $\mathbb{L}_r = (1, \ldots, L_r)$. Similarly, I define the sequence of positions within $\boldsymbol{s}'$ by $\mathbb{L}'_r = (1, \ldots, L')$. Now, we can more precisely state our goal as modeling the posterior distribution $\boldsymbol{B}|\boldsymbol{r}, \boldsymbol{s}', \boldsymbol{c}$. In order to be able to do this, it is convenient to define a matrix-valued random variable $\boldsymbol{\Phi}$ and a vector-valued random variable $\boldsymbol{\Psi}$. The $w \times 4$ matrix random variable $\boldsymbol{\Phi}$ represents the PPM that is associated with the PFM $\boldsymbol{c}$, and it is a random variable because the true probability model will always be uncertain if the number of representative binding site sequences (ie, the number c) is finite. In keeping with a PPM model, for each sample $\boldsymbol{\phi}$ from the random variable $\boldsymbol{\Phi}$, each row of $\boldsymbol{\phi}$ (which I denote by $\boldsymbol{\phi}_g$ where $g \in \mathbb{G}$) has unit L^1 norm. This means that each row $\boldsymbol{\Phi}_g$ of $\boldsymbol{\Phi}$ is a random variable whose range is the unit three-simplex $\mathbb{H}^3$. A central assumption that makes a Bayesian analysis of TFBS enrichment tractable is that the $\boldsymbol{\Phi}_g$ are all independent random variables. The $\mathbb{H}^3$-valued random variable $\boldsymbol{\Psi}$ represents the nucleotide frequencies on $\boldsymbol{s}'$, and its distribution is generally very sharply peaked since the sequence $\boldsymbol{s}'$ from which the background model is obtained is usually hundreds to thousands of kilobase pair in length.

In the application of PFM-guided TFBS enrichment analysis, the observations $\boldsymbol{r}$, $\boldsymbol{c}$, and $\boldsymbol{s}'$ are known by definition; however, it is helpful in a Bayesian approach to formally define a generative model in which we can compute the probability of these observations, conditioned on Λ, $\boldsymbol{B}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$. Such a generative model can be more concisely defined in terms of random variables, and thus, I refer to a $\mathbb{D}^{L'}$-valued random variable $\boldsymbol{R}$ for which we have the observed sequence $\boldsymbol{r}$, and a $\mathbb{D}^{L'}$-valued random variable $\boldsymbol{S}'$ for which we have observed s', and a $\{1, \ldots, c\}^{w \times 4}$-valued random variable $\boldsymbol{C}$ for which we have observed $\boldsymbol{c}$. The random variables in this model are summarized in Table 1.

In order to be able to model the conditional probability of the sequence $\boldsymbol{r}$ given a PFM $\boldsymbol{\phi}$, the specified locations of TFBS $\boldsymbol{\beta}$, and the background nucleotide frequency model $\boldsymbol{\psi}$, it is necessary to define a function $\mathcal{U}: \{-1,0,1\}^{L_r} \to \mathcal{P}(\mathbb{L}_r)$ that maps a configuration $\boldsymbol{\beta}$ of binding sites to the set of nucleotide positions within $\mathbb{L}_r$ that the binding sites occupy. Thus, $\mathcal{U}(\boldsymbol{\beta})$ is the *footprint* of the binding sites whose 5′ locations are specified by $\boldsymbol{\beta}$. Let us define the set of all pairs of binding site *footprint* positions and binding site configurations by $\mathbb{U} \subset \mathbb{L}_r \times \mathbb{B}$:

$$\mathbb{U} = \bigcup_{\boldsymbol{\beta} \in \mathbb{B}} \mathcal{U}(\boldsymbol{\beta}) \times \{\boldsymbol{\beta}\}. \tag{5}$$

Given a configuration of binding sites $\boldsymbol{\beta}$, any position $l \in \mathcal{U}(\boldsymbol{\beta})$ within one of the binding sites will correspond to a specific binding site orientation (1 or −1), and this correspondence will be denoted by a mapping $\mathcal{J}: \mathbb{U} \to \{-1,1\}$,

$$(l, \boldsymbol{\beta}) \underset{\mathcal{J}}{\mapsto} \begin{cases} 1 & \text{if } l \text{ is in a forward-orientation binding site} \\ -1 & \text{if } l \text{ is in a reverse-orientation binding site.} \end{cases} \tag{6}$$

Table 1. Random variables in the full probability model for TFBS enrichment analysis.

VARIABLE	SAMPLE/ OBSERVATION	RANGE	MEANING
Λ	λ	$(0, \lambda_{\max}]$	Frequency (per bp) of binding sites within $\boldsymbol{r}$
$\boldsymbol{B}$	$\boldsymbol{\beta}$	$\{-1,0,1\}^{L_r}$	Presence/absence (and orientation) of TFBS
$\boldsymbol{\Phi}$	$\boldsymbol{\phi}$	$(\mathbb{H}^3)^w$	The true PPM of the TF
$\boldsymbol{\Psi}$	$\boldsymbol{\psi}$	$\mathbb{H}^3$	Nucleotide frequencies on non-TFBS DNA
$\boldsymbol{R}$	$\boldsymbol{r}$	$\mathbb{D}^{L_r}$	Gene-vicinal, noncoding sequence
$\boldsymbol{S}'$	$\boldsymbol{s}'$	$\mathbb{D}^{L'}$	Background noncoding sequence
$\boldsymbol{C}$	$\boldsymbol{c}$	$\{1, \ldots, c\}^{w \times 4}$	The PFM for representative binding sites

In the case of a reverse-orientation binding site, the PPM $\boldsymbol{\phi}$ will correspond to the reverse complement of the nucleotide sequence within the binding site, in which case it is convenient to define a conditional complementation function $\mathcal{C}$: $\mathbb{D} \times \{-1,1\} \to \mathbb{D}$ by

$$\mathcal{C}(d, j) = jd + \frac{5}{2}(1 - j), \tag{7}$$

which is the identity on d when $j = 1$ and which complements d when $j = -1$. Similarly, any configuration $\boldsymbol{\beta}$ and any position l within a binding site will correspond to a specific row of the PFM for the TF, depending on the orientation of the binding site; I denote this correspondence by a mapping $\mathcal{G}$: $\mathbb{U} \to \mathbb{G}$,

$$(l, \boldsymbol{\beta}) \underset{\mathcal{G}}{\mapsto} \text{the index of the row of } \boldsymbol{c} \text{ corresponding to position } l \in \mathcal{U}(\boldsymbol{\beta}). \tag{8}$$

Finally, in order to be able to model the joint probability of $\boldsymbol{r}$ and $\boldsymbol{s}$, it will be necessary to count nucleotides of each type (ie, A, C, G, and T) outside of TFBS as well as at different positions within the binding sites of TFX. Outside of TFBS, I represent the nucleotide counts by the 4-vector $\boldsymbol{f}$ whose elements are defined by

$$f_d = |\{l \in \mathbb{L}_r \setminus \mathcal{U}(\boldsymbol{\beta}) \mid r_l = d\}| + |\{l \in \mathbb{L}' \mid s'_l = d\}|, \tag{9}$$

for all $d \in \mathbb{D}$. I represent the position-nucleotide counts for the sequence within all TFBS by a $w \times 4$ matrix $\boldsymbol{\sigma}$ whose elements are

$$\sigma_{gd} = \left|\left\{ l \in \mathcal{U}(\boldsymbol{\beta}) \mid \mathcal{G}(l, \boldsymbol{\beta}) = g \wedge \mathcal{C}(r_l, \mathcal{J}(l, \boldsymbol{\beta})) = d \right\}\right| \tag{10}$$

for all $g \in \mathbb{G}$ and $d \in \mathbb{D}$. Because of the physical constraint that binding sites for TFX cannot overlap, it follows that

$$\sum_{d \in \mathbb{D}} \sigma_{gd} = \beta, \tag{11}$$

for all $g \in \mathbb{G}$. In the next section, I introduce the statistical approach by defining a joint probability model.

Bayesian Approach to TFBS Enrichment Analysis

Having defined random variables to represent all of the observed information ($\boldsymbol{R}$, $\boldsymbol{C}$, $\boldsymbol{S}'$) and the latent variables ($\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, Λ), and the model parameter $\boldsymbol{B}$, the first step in a Bayesian approach[44] is to define a simplified model for the joint probability distribution. I choose the model

$$p(r, s', \Lambda, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\psi}, c) = P(r \mid \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\beta}) P(s' \mid \boldsymbol{\psi})\, p(\boldsymbol{\psi})\, P(c \mid \boldsymbol{\phi})\, p(\boldsymbol{\phi}) P(\boldsymbol{\beta} \mid \lambda)\, p(\lambda), \tag{12}$$

where the condensed notation $P(\lambda)$ means $P(\Lambda = \lambda)$ and so forth for the other random variables, P denotes a probability distribution, and p denotes a probability density. Eq. 12 can be derived from first principles based on the following independence assumptions:

$$\boldsymbol{R} \perp \boldsymbol{S}', \boldsymbol{C}, \Lambda \mid \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\beta} \tag{13}$$

$$\boldsymbol{S}' \perp \boldsymbol{\Phi}, \boldsymbol{B}, \boldsymbol{C}, \Lambda \mid \boldsymbol{\psi} \tag{14}$$

$$\boldsymbol{C} \perp \Lambda, \boldsymbol{B}, \boldsymbol{\Psi} \mid \lambda \tag{15}$$

$$\boldsymbol{B} \perp \boldsymbol{\Phi}, \boldsymbol{\Psi} \mid \lambda \tag{16}$$

$$\boldsymbol{\Phi} \perp \boldsymbol{\Psi}, \Lambda \tag{17}$$

$$\boldsymbol{\Psi} \perp \Lambda \tag{18}$$

The independence structure of Eq. 12 can be summarized in graphical model notation,[45] as shown in Figure 2. To make the joint probability model explicit, each of the conditional probabilities in Eq. 12 will be specified below.

Modeling $P(r|\phi,\psi,\beta)P(s'|\psi)p(\psi)$. For PFM-guided computational recognition of TFBS, a fundamental assumption is that the probability model for the counts of nucleotides outside of TFBS is independent of the probability model for the counts of nucleotides within TFBS.[32,46] This means that the conditional probability of $\boldsymbol{r}$ can be expressed as the product of conditional probabilities for the subsequences of $\boldsymbol{r}$ corresponding to $\mathcal{U}(\boldsymbol{\beta})$ (the TFBS) and corresponding to $\mathbb{L}_r \setminus \mathcal{U}(\boldsymbol{\beta})$ (outside the TFBS). Conditioned on $\boldsymbol{\psi}$, the nucleotide probabilities at positions outside of TFBS, which are denoted by the random variables $\{R_l\}_{l \in \mathbb{L}_r \setminus \mathcal{U}(\boldsymbol{\beta})}$ and the random variables $\{S_l\}_{l \in \mathbb{L}'}$, are assumed to satisfy

$$R_l \mid \boldsymbol{\psi} \overset{\text{iid}}{\sim} \text{Categ}(\mathbb{D}, \boldsymbol{\psi}) \tag{19}$$

for any $l \in \mathbb{L}_r \setminus \mathcal{U}(\boldsymbol{\beta})$, and

$$S_l \mid \boldsymbol{\psi} \overset{\text{iid}}{\sim} \text{Categ}(\mathbb{D}, \boldsymbol{\psi}), \tag{20}$$

for any $l \in \mathbb{L}'$ (where iid denotes independent and identically distributed). Conditioned on $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$, the nucleotide sequence

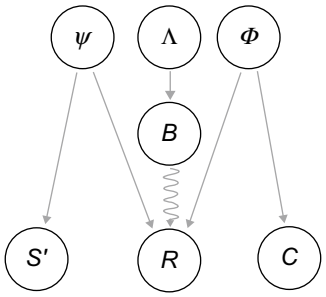

Figure 2. Graphical model diagram of the independence assumptions shown in Eqs. 13–18. Each arrow denotes a relationship between a *parent variable* and a *child variable*. Collectively, the variables and arrows indicate conditional independence as follows: each variable X is independent of other variables, jointly conditioned on all parents of X.

probabilities at locations within the *footprint* $\mathcal{U}(\boldsymbol{\beta})$ of the binding sites specified by $\boldsymbol{\beta}$, the nucleotide probabilities are denoted by random variables $\{R_l\}_{l\in\mathcal{U}(\boldsymbol{\beta})}$ that are independent and distributed as follows:

$$R_l \mid \boldsymbol{\phi},\boldsymbol{\beta},\mathcal{G}(l,\boldsymbol{\beta})=g,\mathcal{J}(l,\beta)=1 \sim \text{Categ}(\mathbb{D},\phi_g) \quad (21)$$

$$5-R_l \mid \boldsymbol{\phi},\boldsymbol{\beta},\mathcal{G}(l,\boldsymbol{\beta})=g,\mathcal{J}(l,\beta)=-1 \sim \text{Categ}(\mathbb{D},\phi_g). \quad (22)$$

for any $l\in\mathbb{L}_r\backslash\mathcal{U}(\boldsymbol{\beta})$. Because the length of $\boldsymbol{s}'$ is assumed to be quite substantial, the distribution of $\boldsymbol{\Psi}|\boldsymbol{s}'$ will be quite sharply peaked, and thus, any weak prior on ψ will have little effect. Thus, it is reasonable to assume a uniform prior $p(\boldsymbol{\psi})=3!$. Given the uniform prior assumption for $\boldsymbol{\Psi}$ and the definition in Eqs. 9 and 10 and the assumptions in Eqs. 19–22, the conditional probability of $\boldsymbol{R}$, $\boldsymbol{S}'$ has a compact form,

$$P(r \mid \boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\beta})P(s' \mid \boldsymbol{\psi})p(\boldsymbol{\psi})=3!\prod_{d\in\mathbb{D}}\left[(\psi_d)^{f_d}\left(\prod_{g\in\mathbb{G}}(\phi_{gd})^{\sigma_{gd}}\right)\right], \quad (23)$$

that will be compatible with collapsing of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ (as shown in the "Obtaining the distribution of $\boldsymbol{B}|\boldsymbol{r},\boldsymbol{s}',\boldsymbol{c}$" section).

Modeling $P(\boldsymbol{c}|\boldsymbol{\phi})\,p(\boldsymbol{\phi})$. To account for uncertainty in the PFM $\boldsymbol{c}$ due to sampling from a finite (and in many cases, very limited) number of representative binding site sequences, the PFM is represented by a random variable $\boldsymbol{C}$. A core assumption in the field of PFM-guided TFBS recognition is that rows of $\boldsymbol{C}$, denoted by $\boldsymbol{C}_g$ (where $g\in\mathbb{G}$), are independent and multinomial distributed with a fixed number of trials.[47,48] Because in some cases, some representative binding site sequences will be outside the core portion of the multiple alignment from which the PFM is tabulated, the row sums of $\boldsymbol{c}$ may in some cases be less than the count c of representative binding site sequences. Thus, to accommodate such cases, I denote by c_g the sum of the elements of row g of $\boldsymbol{c}$. In terms of the row-specific counts c_g (for $g\in\mathbb{G}$), the conditional distribution of $\boldsymbol{C}_g$ can be expressed as

$$\boldsymbol{C}_g \,\Big|\, \|\boldsymbol{C}_g\|_1=c_g,\boldsymbol{\Phi}_g=\boldsymbol{\phi}_g \sim \text{Mult}(c_g,\boldsymbol{\phi}_g), \quad (24)$$

for which the formula for $P(\boldsymbol{c}|\boldsymbol{\phi})$ immediately follows

$$P(c \mid \boldsymbol{\phi})=\prod_{g\in\mathbb{G}}\left[\frac{c_g!}{\prod_{d\in\mathbb{D}}c_{gd}!}\prod_{d'\in\mathbb{D}}(\phi_{gd'})^{c_{gd'}}\right]. \quad (25)$$

The most common approach for selecting the prior probability $p(\boldsymbol{\phi})$ for the PPM is to choose a uniform prior, in which case $p(\boldsymbol{\phi})$ is just the constant $(3!)^w$. Although other authors have pointed out the possibility of using an empirical prior on $\boldsymbol{\phi}$,[30] it is nontrivial to collapse $\boldsymbol{\Phi}$ by analytic integration over all $\boldsymbol{\phi}$, in the case of a nonuniform $p(\boldsymbol{\phi})$, so here I assume a uniform $p(\boldsymbol{\phi})$.

Modeling $P(\boldsymbol{\beta}|\lambda)p(\lambda)$. At a given base pair location with no binding sites nearby (and with no sequence information), I model the probability that there is a binding site – in a specific orientation – as $\lambda/2$. In the absence of sequence information, intuition would suggest treating the occurrence or absence of a binding site for TFX at each position in DNA as independent and identically distributed Bernoulli trials. However, because of the physical constraint that two binding sites are not allowed to overlap, each binding site (ie, each nonzero entry of $\boldsymbol{\beta}$) affects the probability of a binding site at nearby positions. Specifically, each binding site prevents the possibility of an overlapping binding site (in either orientation) at $w-1$ bp positions, and for an additional $2(w-1)$ flanking positions, a binding site is only possible in one orientation. Thus, the probability model consistent with the physical constraints would be

$$P(\boldsymbol{\beta} \mid \lambda)=\mathcal{N}_1(L_r,w,\lambda)(\lambda/2)^{\beta}(1-\lambda)^{L_r-w\beta-2(w-1)\beta}\times\left(1-\frac{\lambda}{2}\right)^{2(w-1)\beta}, \quad (26)$$

where $\mathcal{N}_1$ is function that is implicitly defined by the law of total probability for $P(\boldsymbol{\beta}|\lambda)$. In the limit where $L_r\gg w$, and solving for $\mathcal{N}_1$ using the law of total probability, we have the approximate result,

$$P(\boldsymbol{\beta} \mid \lambda)\simeq\left(1+\frac{\lambda w\left(1-\frac{\lambda}{2}\right)^{2(w-1)}}{(1-\lambda)^{3w-2}}\right)^{-\frac{L_r}{w}}\cdot\left(\frac{\lambda\left(1-\frac{\lambda}{2}\right)^{2(w-1)}}{2(1-\lambda)^{3w-2}}\right)^{\beta}+O(w/L_r). \quad (27)$$

In the case $w=1$, the above can be seen to reduce to $\lambda^{\beta}(1-\lambda)^{L_r-\beta}/2^{\beta}$, which is the expected joint probability of outcome sequence $\boldsymbol{\beta}$ for L_r independent trials of the categorical distribution with outcomes $(-1,0,1)$ with probabilities $(\lambda/2,1-\lambda,\lambda/2)$, in which $\boldsymbol{\beta}$ trials have a nonzero outcome.

The prior distribution $p(\lambda)$ reflects the range and relative probability of different Λ values for TFX, before the sequence $\boldsymbol{r}$ has been taken into account. The prior $p(\lambda)$ is important because for real-world applications, it can exert a significant effect on the distribution of $\Lambda|\boldsymbol{r},\boldsymbol{c}$. For mammals, the prior $p(\lambda)$ can be formulated empirically using binding site frequencies (per base pair of noncoding, gene-vicinal DNA sequence) for 620 human TF ChIP-seq experiments (comprising 119 distinct TFs) obtained from the ENCODE project.[43] For each ChIP-seq experiment, binding sites within regions of noncoding DNA within –1500 to +500 bp transcription start sites of VEGA transcripts (from Ensembl Release 75, GRCh37 assembly coordinates) were

mapped, using ChIP-seq peak data that were peak-called using the SPP program[49] and for which the data files were downloaded from the ENCODE data access page at the European Bioinformatics Institute from the June 2012 release (http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_datajan2011/byDataType/peaks/june2012/spp/optimal/) in narrowPeak format. The counts of binding sites within these noncoding regions were fit to a Poisson model parameterized by a binding site frequency λ per base pair of DNA; for each ChIP-seq experiment, a λ estimate was obtained using maximum likelihood. The resulting histogram of λ estimates is well-described by a beta distribution, as shown in Figure 3, with parameters as given in Table 2. Thus, it is convenient to adopt a prior density.

$$p(\lambda) = \frac{1}{\mathcal{B}(\lambda_{max};\alpha,\nu)}\lambda^{\alpha-1}(1-\lambda)^{\nu-1}, \tag{28}$$

where $\mathcal{B}$ is the incomplete beta function, and the shape hyperparameters are as given in Table 2 (recall that the range of Λ is $(0,\lambda_{max}]$). Combining Eqs. 27 and 28, and in the limit where $\lambda \ll \sqrt{L_r/w}$, the product $P(\boldsymbol{\beta}|\lambda)p(\lambda)$ can be approximated

$$P(\boldsymbol{\beta} \mid \lambda)p(\lambda) = \frac{1}{2^{\beta}\mathcal{B}(\lambda_{max};\alpha,\nu)}\lambda^{\beta+\alpha-1}(1-\lambda)^{\nu+L_r-(2w-1)\beta-1} \times (1-O(\lambda^2)), \tag{29}$$

where with our choice of λ_{max}, the second-order term can be neglected, resulting in a beta distribution-like dependence on λ,

$$P(\boldsymbol{\beta} \mid \lambda)p(\lambda) = p(\boldsymbol{\beta},\lambda) = \frac{1}{2^{\beta}\mathcal{B}(\lambda_{max};\alpha,\nu)}\lambda^{\beta+\alpha-1} \times (1-\lambda)^{\nu+L_r-(2w-1)\beta-1}, \tag{30}$$

for $\lambda \in (0,\lambda_{max}]$. From this equation, and integrating λ over the range from $[0,\lambda_{max}]$, we see that the probability model Eq. 26 corresponds to a prior

$$P(\boldsymbol{\beta}) = \frac{\mathcal{B}(\lambda_{max};\beta+\alpha,\nu+L_r-(2w-1)\beta)}{2^{\beta}\mathcal{B}(\lambda_{max};\alpha,\nu)}. \tag{31}$$

In contrast to Eq. 31, Lähdesmäki and Shmulevich[50] used a geometric prior on β, $P(\beta) = 1/2^{\beta+1}$, implying

$$P(\boldsymbol{\beta}) = q\frac{\beta!(L_r-\beta)!}{L_r!}\frac{1}{2^{\beta+1}}, \tag{32}$$

up to a normalization constant q. A comparison of the two priors (Fig. 4) suggests that the incomplete beta function prior (Eq. 31) may be more conservative than the geometric prior – a hypothesis that I investigate by analyzing simulated promoter sequence data in the next section.

Table 2. Parameter estimate for the beta distribution model for the prior $p(\lambda)$ on the binding site frequency per base pair, for human transcription factors.

	α	ν ($\times 10^4$)
Least-squares estimate	1.37	3.62
95% confidence interval	±0.15	±0.53

Obtaining the Distribution of $B|r, s', c$

The second step in a Bayesian approach[44] is to condition on the observed data – in this case, $\boldsymbol{r}$, $\boldsymbol{s'}$, and $\boldsymbol{c}$ – and then obtain the conditional distribution of the parameter(s) of interest, in this case $\boldsymbol{B}$. In order to be able to do so, starting from the joint probability model (Eq. 12), the nuisance parameters $\boldsymbol{\phi}$, $\boldsymbol{\psi}$, and λ must be either estimated or marginalized. A key advantage of the Bayesian approach is that we can take into account the probability distribution of $\Phi|\boldsymbol{c}$, in the process of eliminating Φ by marginalization. The parameter ψ can be similarly marginalized, although the uncertainty in the background nucleotide

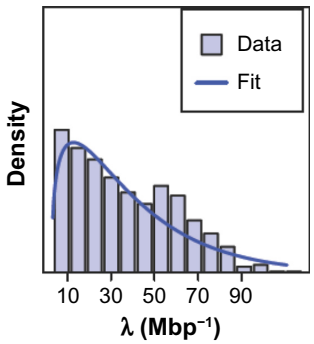

Figure 3. Distribution of frequencies of TFBS (per base pair of noncoding, gene-vicinal DNA sequence) for human transcription factors, based on the analysis of 620 ChIP-seq datasets from the ENCODE project.[43]

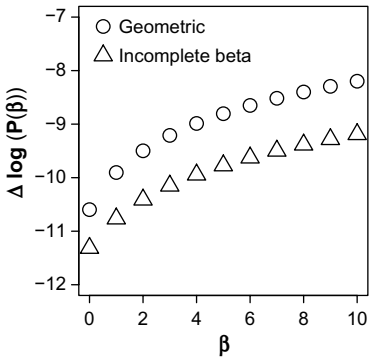

Figure 4. Plot of the change in log $P(\boldsymbol{\beta})$ with the addition of a single binding site, as a function of β, for the incomplete beta function-based prior (Eq. 31) and the previously proposed geometric prior (Eq. 32). The Δ log P values between the two priors are closer for $\beta = 0$ but become greater with increasing β, indicating that the empirical prior in Eq. 31 is not simply equivalent to a rescaling of the geometric prior.

frequency is generally very small for real-world applications in which L' is large. We marginalize $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ by integration,

$$p(\boldsymbol{r},\boldsymbol{s}',\lambda,\boldsymbol{\beta},\boldsymbol{c}) = \int \mathrm{d}^{3w}\boldsymbol{\phi} \int \mathrm{d}^{3}\boldsymbol{\psi}\ p(\boldsymbol{r},\boldsymbol{s}',\lambda,\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{c}). \tag{33}$$

Given Eqs. 12, 23, and 25, the dependence of the integrand in Eq. 33 on $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ has the same algebraic form as the probability density function for independent Dirichlet random variables $\{\Psi, \Phi_1, \ldots, \Phi_w\}$,as a consequence of the fact that the Dirichlet distribution is the conjugate prior for the multinomial distribution.[44] Thus, the two integrals in Eq. 33 can be evaluated analytically.[30] The desired joint conditional probability of Λ, $\boldsymbol{B}$ follows by the definition of conditional probability,

$$p(\lambda,\boldsymbol{\beta} \mid \boldsymbol{r},\boldsymbol{s}',\boldsymbol{c}) = \frac{p(\boldsymbol{r},\boldsymbol{s}',\lambda,\boldsymbol{\beta},\boldsymbol{c})}{P(\boldsymbol{r},\boldsymbol{s}',\boldsymbol{c})}. \tag{34}$$

After performing the integrals in Eq. 33, using Eqs. 30 and 34, and then log-transforming, we have

$$\begin{aligned}
&\log p(\lambda,\boldsymbol{\beta} \mid \boldsymbol{r},\boldsymbol{s}',\boldsymbol{c}) \\
&= \log \mathcal{N}_2(\lambda_{\max}, L_r, w, \alpha, \nu, \boldsymbol{c}) - \log P(\boldsymbol{r},\boldsymbol{s}',\boldsymbol{c}) - \beta \log 2 \\
&\quad + (\alpha+\beta-1)\log\lambda + (\nu + L_r - (2w-1)\beta - 1)\log(1-\lambda) \\
&\quad + \sum_{g\in\mathbb{G}}\sum_{d\in\mathbb{D}} \log(c_{gd}+\sigma_{gd})! - \sum_{g\in\mathbb{G}} \log(c_g+\beta+3)! \\
&\quad + \sum_{d\in\mathbb{D}} \log f_d! - \log(L_r + L' - \beta w + 3)!,
\end{aligned} \tag{35}$$

where $\mathcal{N}_2(\lambda_{\max}, L_r, w, \alpha, \nu, \boldsymbol{c})$ is function that will not need to be evaluated. The parameter λ can then be marginalized by integration, yielding

$$\begin{aligned}
&\log P(\boldsymbol{\beta} \mid \boldsymbol{r},\boldsymbol{s}',\boldsymbol{c}) \\
&= \log \mathcal{N}_2(\lambda_{\max}, L_r, w, \alpha, \nu, c) - \log P(\boldsymbol{r},\boldsymbol{s}',\boldsymbol{c}) - \beta \log 2 \\
&\quad + \log \mathcal{B}(\lambda_{\max}; \alpha+\beta, \nu + L_r - (2w-1)\beta) \\
&\quad + \sum_{g\in\mathbb{G}}\sum_{d\in\mathbb{D}} \log(c_{gd}+\sigma_{gd})! - \sum_{g\in\mathbb{G}} \log(c_g+\beta+3)! \\
&\quad + \sum_{d\in\mathbb{D}} \log f_d! - \log(L_r + L' - \beta w + 3)!,
\end{aligned} \tag{36}$$

where $\mathcal{B}$ is the incomplete beta function. As we will see below, in order to obtain samples of $\boldsymbol{B}|\boldsymbol{r}, \boldsymbol{s}', \boldsymbol{c}$, it will not be necessary to explicitly evaluate $P(\boldsymbol{r}, \boldsymbol{s}', \boldsymbol{c})$. Now that we have an explicit formula for $\log[P(\boldsymbol{\beta}|\boldsymbol{r}, \boldsymbol{s}', \boldsymbol{c})]$ up to additive terms that do not depend on $\boldsymbol{\beta}$, it is possible to generate $\boldsymbol{\beta}$ samples from this distribution using Markov Chain Monte Carlo (MCMC) sampling.

MCMC approach. For sampling from $\boldsymbol{B}|\boldsymbol{r}, \boldsymbol{s}', \boldsymbol{c}$, the Metropolis–Hastings algorithm,[51] in which a probabilistic proposal generator $g(\boldsymbol{\beta}, \boldsymbol{\beta}')$ for a transition from $\boldsymbol{\beta} \to \boldsymbol{\beta}'$ can be defined so as to optimize the acceptance rate for moves, is convenient. For the problem of TFBS enrichment detection, following the general approach used by Lähdesmäki et al for TFBS recognition, I use a two-stage proposal generator in which a base pair position $l \in \mathbb{L}_r$ is selected at random, and then, depending on the current state of $\boldsymbol{\beta}$, binding site removal or addition (in the latter case, with a randomly selected value $j \in \{-1,1\}$) is proposed (in the case of binding site addition, $j = -1$ or $j = 1$ is chosen with equal probability). For this approach, it will be useful to have a simplified expression for the log probability ratio for $B_l = j$ versus $B_l = 0$, conditioned on $\boldsymbol{r}, \boldsymbol{s}', \boldsymbol{\beta}_{\mathbb{L}_r\setminus\{l\}}, \boldsymbol{c}$; it is convenient to define some additional notation in order to make this conditional probability ratio explicit. Without loss of generality, let us assume that the current state for the *hypothetical* binding site configuration $\boldsymbol{\beta} \in \mathbb{B}$, a location $l \in \mathbb{L}_r$ such that $\beta_l = 0$, and an orientation $j \in \{-1,1\}$ such that the configuration $\boldsymbol{\beta}$ with $\beta_l = j$ would not violate the TFBS physical constraints. In order to simplify notation, I define a function $\mathcal{H}$: $\mathbb{D}^{L_r} \times \mathbb{L} \times \{-1,1\} \times \mathbb{G} \to \mathbb{D}$ by

$$\mathcal{H}(\boldsymbol{r}, l, j, g) = \mathcal{C}(r_{l+j(g-1)}, j), \tag{37}$$

whose value represents the nucleotide at position g within the binding site for TFX that has orientation j and whose 5′ most nucleotide is at location $l \in \mathbb{L}_r$. I also define a function $\mathcal{I}$: $\mathbb{D}^{L_r} \times \mathbb{L}_r \times \{-1,1\} \times \mathbb{D} \to \{0, 1, \ldots, g\}$ by

$$\mathcal{I}(\boldsymbol{r}, l, j, d) = \left| \left\{ g \in \mathbb{G} \mid d = r_{l+j(g-1)} \right\} \right|, \tag{38}$$

whose value is the count of nucleotides of base d within a binding site for TFX in orientation j whose 5′ most nucleotide is at location l.

Applying Eq. 35 to two different binding site configurations that differ by one binding site being present/absent at a specific location $l \in \mathbb{L}_r$, and using the definitions of $\mathcal{H}$ and $\mathcal{I}$, we obtain a closed-form expression for the log ratio of the conditional probability of there being a binding site at location l (in orientation j), to the conditional probability that there is not a binding site at l

$$\begin{aligned}
&\log\left(\frac{P(B_l = j \mid \boldsymbol{r},\boldsymbol{s}',\boldsymbol{\beta}_{\mathbb{L}_r\setminus\{l\}},\boldsymbol{c})}{P(B_l = 0 \mid \boldsymbol{r},\boldsymbol{s}',\boldsymbol{\beta}_{\mathbb{L}_r\setminus\{l\}},\boldsymbol{c})} \right) \\
&= -\log 2 \\
&\quad + \log \mathcal{B}(\lambda_{\max}; \alpha+\beta+1, \nu + L_r - (2w-1)(\beta+1)) \\
&\quad - \log \mathcal{B}(\lambda_{\max}; \alpha+\beta, \nu + L_r - (2w-1)\beta) \\
&\quad + \sum_{g\in\mathbb{G}} \log(c_{g,\mathcal{H}(r,l,j,g)} + \sigma_{g,\mathcal{H}(r,l,j,g)} + 1) \\
&\quad - \sum_{g\in\mathbb{G}} \log(c_g + \beta + 4) \\
&\quad - \sum_{d\in\mathbb{D}} \log\left\{ (f_d)_{[\mathcal{I}(r,l,j,d)]} \right\} \\
&\quad + \log\left\{ (L_r + L' - \beta w + 3)_{[w]} \right\},
\end{aligned} \tag{39}$$

where the notation $(x)_{[n]}$ denotes the falling factorial and $\boldsymbol{\sigma}$ is computed for $\boldsymbol{\beta}$. Given Eq. 39, sampling from the distribution of $B|r, s', c$ can be accomplished using the Metropolis–Hastings algorithm with the following proposal distribution:

$$g\left((\beta_1,\ldots,\beta_{l-1},j,\beta_{l+1},\ldots,\beta_{L_r}) \mid (\beta_1,\ldots,\beta_{l-1},0,\beta_{l+1},\ldots,\beta_{L_r})\right) = \frac{(S_l)^q}{\sum_{l' \in \mathbb{L}_r} (S_{l'})^q}, \quad (40)$$

where S_l is the Shannon entropy of the probabilities $P(\beta_l = j | \boldsymbol{r}, \boldsymbol{s'}, \boldsymbol{\beta}_{\mathbb{L}_r \setminus \{l\}}, \boldsymbol{c})$ (for $j \in \{-1,0,1\}$) with $\boldsymbol{\beta}_{\mathbb{L}_r \setminus \{l\}} = \mathbf{0}$, and where the weight exponent q is tuned to achieve the desired average acceptance rate.[52] The reason for using a proposal distribution that weights each position by the Shannon entropy is that at most positions $l \in \mathbb{L}_r$, the entropy of the three conditional probabilities $P(\beta_l = j \mid \boldsymbol{r}, \boldsymbol{s'}, \boldsymbol{\beta}_{\mathbb{L}_r \setminus \{l\}}, \boldsymbol{c})$ (for $j \in \{-1,0,1\}$ and fixed l) is very small, and thus, from the standpoint of optimizing the acceptance rate, it is convenient to weight the generation of proposed moves toward moves with more even odds of move acceptance. Results from empirical testing with sequence lengths $L_r = 2 \times 10^4$ suggest that a value $q = 0.85$ gives an acceptance rate of about 0.12, with increasing values of q increasing the acceptance rate. In this study, the Markov chain is initialized by iterating over $l = 0$ to $l = L_r$, for each l, setting β_l to be the most probable configuration given Eq. 39, conditioned on $\boldsymbol{\beta}_{\mathbb{L}_r \setminus \{l\}} = \mathbf{0}$. Once the Markov chain has converged, at most positions $l \in \mathbb{L}_r$, the entropy of the three conditional probabilities $P(\beta_l = j \mid \boldsymbol{r}, \boldsymbol{s'}, \boldsymbol{\beta}_{\mathbb{L}_r \setminus \{l\}}, \boldsymbol{c})$ (for $j \in \{-1,0,1\}$ and fixed l) is very small, and thus, from the standpoint of optimizing the acceptance rate, it is convenient to weight the generation of proposed moves toward moves with more even odds of move acceptance.

Empirical Results

Based on a direct comparison (Fig. 4) of the incomplete beta function prior (Eq. 31) and the previously proposed geometric prior (Eq. 32), it seems reasonable to suppose that, within the context of a Bayesian approach for PFM-based TFBS frequency estimation (as described in the "Obtaining the distribution of $\boldsymbol{B}|\boldsymbol{r}, \boldsymbol{s'}, \boldsymbol{c}$" section) the incomplete beta function prior and the geometric prior might have different effects on the conditional distribution of the number of TFBS, ie, the samples of $\boldsymbol{B}|\boldsymbol{c}, \boldsymbol{r}, \boldsymbol{s'}$. To test this hypothesis, I generated a synthetic dataset based on a simulated background sequence $\boldsymbol{s'}$ (with $L' = 100{,}000$) and 120 gene promoter sequences (each with $L_r = 20{,}000$), with uniform probabilities for each nucleotide. Into each simulated base sequence $\boldsymbol{r}$, and for each of a fixed set of 100 TF PFMs selected at random from TRANSFAC Professional 2015, $t \in \{1, \ldots, 10\}$ TFBS were inserted into the $\boldsymbol{r}$ sequence (using representative binding site sequences from which the PFMs were computed, resulting in a modified sequence $\boldsymbol{r}_t$). Ten samples from the stationary distribution of B (the number of TFBS) were then generated using the MCMC approach described in the "MCMC approach" section (with 5000 burn-in steps, 100 steps per sample, and $q = 0.85$), for both the geometric prior and the incomplete beta function-based prior (with $\nu = 10{,}000$ and $\alpha = 1.0$). For each of the two priors and for each combination of sequence $\boldsymbol{r}$, PFM $\boldsymbol{c}$, and number of ground-truth binding sites t, the 10 $\boldsymbol{B}|\boldsymbol{c}, \boldsymbol{r}_t, \boldsymbol{s'}$ samples were averaged, producing one *geometric prior* sample and one *incomplete beta function prior* sample for each of the 120,000 combinations of $\boldsymbol{c}$, $\boldsymbol{r}$, and t. The distributions of $\boldsymbol{B}|\boldsymbol{c}, \boldsymbol{r}_t, \boldsymbol{s'}$, organized by t and by prior, as shown in Figure 5, reveal several interesting patterns. First, across the fixed set of 100 randomly selected TFs, the MCMC method incorporating the incomplete beta function prior appears to yield samples that are more accurate than the MCMC method incorporating the geometric prior. In terms of mean-squared error, the MCMC method with the incomplete beta function-based prior is 19.6, whereas the mean-squared error with the geometric prior is 107.7. Second, the samples generated using the MCMC method with the geometric prior appear to be substantially higher-variance than the samples generated using the MCMC method with the incomplete beta function-based prior (quantitatively, the t-averaged standard deviation of the TFBS count samples obtained using the incomplete beta function prior was 4.05 versus 8.75 for the geometric prior.

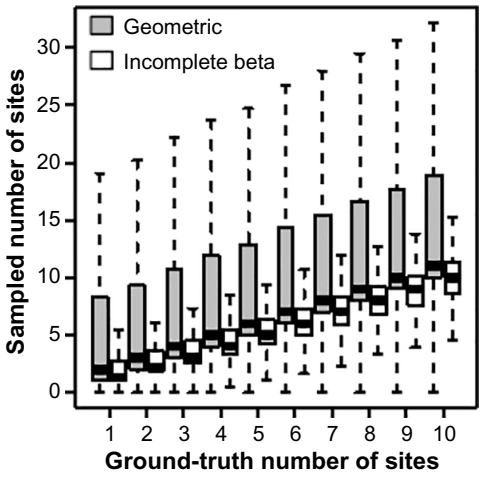

Figure 5. Comparing the accuracies of two MCMC implementations of the Bayesian method for estimating the number of binding sites of a TF, based on the geometric prior (Eq. 32) and the incomplete beta function-based prior (Eq. 31). For each combination of t (number of ground-truth sites) and type of prior, the bar denotes the median, the box denotes the interquartile range, and the whiskers are offset 1.5 interquartile range above or below the 75th and 25th percentiles.

Discussion

This study demonstrates the utility of incorporating an empirical prior on the TFBS frequency per base pair within the context of a Bayesian method for PFM-based TFBS enrichment analysis, but there are several aspects in which the work raises interesting questions that could be explored in future studies. First, in this work, a two-parameter

parametric function has been fit to empirical data on the density distribution of frequencies of human TFBS per base pair of noncoding, gene-vicinal sequence. Thus, the results shown here do not reveal to what extent the estimated parameters for the distribution would generalize to TFBS frequencies in other species. At least for the mouse genome, available evidence from the modENCODE project suggests that overall, TF binding within promoter regions is highly conserved between human and mouse.[53] Moreover, for two TFs whose TFBS were assayed in five mammalian species by ChIP-seq, the numbers of genome-wide binding sites did not vary more than 2× between species.[22] Thus, it seems reasonable to expect that the Λ prior distribution (across TFs) would be similar, for gene-vicinal noncoding sequence. Nevertheless, in future work, it would be informative to estimate the hyperpriors α and ν for human, mouse, fruit fly, and worm to enable a cross-species comparison. Second, it would be useful to characterize how the choice of q parameter affects the empirical performance of the MCMC approach used here, ie, the acceptance ratio, the number of steps required for burn-in, and the number of steps required between samples; it may be possible to significantly improve the speed of the proposed MCMC method through tuning q and the sampling parameters. Third, a key aspect to be explored is the extent to which the accuracy improvement with the incomplete beta function-dependent prior is associated with high-count versus low-count PFMs. Intuitively, it seems reasonable to suppose that for most TFs, an increase in the accuracy of the prior would be expected to have more of an effect on the posterior distribution of β when the PFM count is low, since a higher count PFM would be expected to have a much bigger likelihood ratio that would, in turn, be more likely to dominate over the prior on the number of TFBS.

Conclusions

This study presents a Bayesian approach to the bioinformatics problem of PFM-guided TFBS enrichment analysis. The method incorporates an empirical prior on the frequency distribution λ of binding sites for TFs that is based on genome location data from the ENCODE project. In addition, the method incorporates a probabilistic model for TFBS occurrence conditioned on the parameter λ that takes into account the finite width of the TFBS, in contrast to a previous approach in which the TFBS probability was assumed to have a geometric dependence with a fixed factor of 1/2.[30] The sampling equation for adding/removing a binding site (Eq. 39) could be easily extended to include other sources of information, such as a regulatory potential score derived from phylogenetic sequence conservation or from epigenetic measurements. The R software code implementing the MCMC method described in the "MCMC approach" section and the promoter analyses shown in Figure 5 is available at http://github.com/ramseylab/tfbsincbeta.

Acknowledgments

The author thanks Jichen Yang, Tanjin Xu, Holly Arnold, and Theo Knijnenburg, for reviewing early drafts of the article and for providing helpful feedback. The auhor also thanks Harri Lähdesmäki for providing technical insights on the Lähdesmäki–Rust–Shmulevich method for TFBS recognition and Yuan Jiang for technical advice. Part of this work was carried out in the laboratories of Alan Aderem and Ilya Shmulevich, and their support is gratefully acknowledged.

Author Contributions

Conceived and designed the experiments: SAR. Analyzed the data: SAR. Wrote the first draft of the manuscript: SAR. Contributed to the writing of the manuscript: SAR. Agree with manuscript results and conclusions: SAR. Jointly developed the structure and arguments for the paper: SAR. Made critical revisions and approved final version: SAR. The author reviewed and approved of the final manuscript.

REFERENCES

1. Stormo GD. Computer methods for analyzing sequence recognition of nucleic acids. *Annu Rev Biophys Biophys Chem*. 1988;17:241–63.
2. Roth FP, Hughes JD, Estep PW, Church GM. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol*. 1998;16:939–45.
3. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*. 2004;5:276–87.
4. Rahmann S, Müller T, Vingron M. On the power of profiles for transcription factor binding site detection. *Stat Appl Genet Mol Biol*. 2003;2:Article7.
5. Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*. 2003;31:3576–9.
6. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*. 2003;31:1753–64.
7. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res*. 2004;32:1372–81.
8. Tompa M, Li N, Bailey TL, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*. 2005;23:137–44.
9. Gilchrist M, Thorsson V, Li B, et al. Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature*. 2006;441:173–8.
10. Tan K, Tegner J, Ravasi T. Integrated approaches to uncovering transcription regulatory networks in mammalian cells. *Genomics*. 2008;91:219–31.
11. Ramsey SA, Klemm SL, Zak DE, et al. Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics PLOS. *Comput Biol*. 2008;4:e1000021.
12. Litvak V, Ratushny AV, Lampano AE, et al. A FOXO3-IRF7 gene regulatory circuit limits inflammatory sequelae of antiviral responses. *Nature*. 2012;490:421–5.
13. Gold ES, Ramsey SA, Sartain MJ, et al. ATF3 protects against atherosclerosis by suppressing 25-hydroxycholesterol-induced lipid body formation. *J Exp Med*. 2012;209:807–17.
14. Ramsey SA, Vengrenyuk Y, Menon P, et al. Epigenome-guided analysis of the transcriptome of plaque macrophages during atherosclerosis regression reveals activation of the Wnt signaling pathway. *PLoS Genet*. 2014;10:e1004828.
15. Quandt K, Frech K, Karas H, Wingender E, Werner T. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res*. 1995;23:4878–84.
16. Wingender E, Chen X, Hehl R, et al. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*. 2000;28:316–9.
17. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*. 2004;32:D91–4.
18. Newburger DE, Bulyk ML. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res*. 2009;37:D77–82.

19. Wang J, Zhuang J, Iyer S, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res*. 2012;22:1798–812.
20. Sebastian A, Contreras-Moreira B. footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*. 2014;30:258–65.
21. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science*. 2007;316:1497–502.
22. Schmidt D, Wilson MD, Ballester B, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*. 2010;328:1036–40.
23. Villar D, Flicek P, Odom DT. Evolution of transcription factor binding in metazoans—mechanisms and functional implications. *Nat Rev Genet*. 2014;15:221–33.
24. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*. 2009;10:252–63.
25. Ho Sui SJ, Mortimer JR, Arenillas DJ, et al. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res*. 2005;33:3154–64.
26. Sinha S. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*. 2006;22:e454–63.
27. Pavesi G, Zambelli F. Prediction of over represented transcription factor binding sites in co-regulated genes using whole genome matching statistics. In: Masulli F, Mitra S, Pasi G, eds. *Applications of Fuzzy Sets Theory*. Berlin: Springer; 2007:651–8.
28. Cheng C, Alexander R, Min R, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res*. 2012;22:1658–67.
29. Sinha S, Tompa M. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*. 2003;31:3586–8.
30. Lähdesmäki H, Rust AG, Shmulevich I. Probabilistic inference of transcription factor binding from multiple data sources. *PLoS One*. 2008;3:e1820.
31. Miller AK, Print CG, Nielsen PMF, Crampin EJ. A Bayesian search for transcriptional motifs. *PLoS One*. 2010;5:e13897.
32. Berg OG. Selection of DNA binding sites by regulatory proteins: the LexA protein and the arginine repressor use different strategies for functional specificity. *Nucleic Acids Res*. 1988;16:5089–105.
33. Nishida K, Frith MC, Nakai K. Pseudocounts for transcription factor binding sites. *Nucleic Acids Res*. 2009;37:939–44.
34. Liu JS. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J Am Stat Assoc*. 1994;89:958–66.
35. Thijs G, Marchal K, Lescot M, et al. A Gibbs sampling method to detect over-represented motifs in the upstream regions of coexpressed genes. *J Comput Biol*. 2002;9:447–64.
36. Xing EP, Wu W, Jordan MI, Karp RM. LOGOS: a modular Bayesian model for de novo motif detection. *Proc IEEE Comput Soc Bioinform Conf*. 2003;2:266–76.
37. Jensen ST, Liu XS, Zhou Q, Liu JS. Computational discovery of gene regulatory binding motifs: a Bayesian perspective. *Stat Sci*. 2004;19:188–204.
38. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005;15:1034–50.
39. Taylor J, Tyekucheva S, King DC, Hardison RC, Miller W, Chiaromonte F. ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res*. 2006;16:1596–604.
40. Berger MF, Badis G, Gehrke AR, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*. 2008;133:1266–76.
41. Gabdoulline R, Eckweiler D, Kel A, Stegmaier P. 3DTF: a web server for predicting transcription factor PWMs using 3D structure-based energy calculations. *Nucleic Acids Res*. 2012;40:W180–5.
42. Muratani M, Deng N, Ooi WF, et al. Nanoscale chromatin profiling of gastric adenocarcinoma reveals cancer-associated cryptic promoters and somatically acquired regulatory elements. *Nat Comm*. 2014;5:4361.
43. Gerstein MB, Kundaje A, Hariharan M, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*. 2012;489:91–100.
44. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. 3rd ed. CRC Press, Boca Raton, FL; 2013.
45. Buntine WL. Operations for learning with graphical models. *J Artif Intel Res*. 1994;2:159–225.
46. Heumann JM, Lapedes AS, Stormo GD. Neural networks for determining protein specificity and multiple alignment of binding sites. *Proc Int Conf Intel Syst Mol Biol*. 1994;2:188–94.
47. Hertz GZ, Hartzell GW, Stormo GD. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput App Biosci CABIOS*. 1990;6:81–92.
48. Jensen ST, Liu JS. Bayesian clustering of transcription factor binding motifs. *J Am Stat Assoc*. 2008;103:188–200.
49. Kharchenko P, Tolstorukov M, Park P. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*. 2008;26(12):1351–9.
50. Lähdesmäki H, Shmulevich I. Learning the structure of dynamic Bayesian networks from time series and steady state measurements. *Mach Learn*. 2008;71:185–217.
51. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys*. 1953;21:1087–92.
52. Roberts GO, Gelman A, Gilks WR. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann Appl Probab*. 1997;7:110–20.
53. Shen Y, Yue F, McCleary DF, et al. A map of the cis-regulatory sequences in the mouse genome. *Nature*. 2012;488:116–20.