OPEN ACCESS Freely available online

PLOS ONE

# Bayesian Centroid Estimation for Motif Discovery

**Luis Carvalho***

Department of Mathematics and Statistics, Boston University, Boston, Massachusetts, United States of America

## Abstract

Biological sequences may contain patterns that signal important biomolecular functions; a classical example is regulation of gene expression by transcription factors that bind to specific patterns in genomic promoter regions. In motif discovery we are given a set of sequences that share a common motif and aim to identify not only the motif composition, but also the binding sites in each sequence of the set. We propose a new centroid estimator that arises from a refined and meaningful loss function for binding site inference. We discuss the main advantages of centroid estimation for motif discovery, including computational convenience, and how its principled derivation offers further insights about the posterior distribution of binding site configurations. We also illustrate, using simulated and real datasets, that the centroid estimator can differ from the traditional maximum a posteriori or maximum likelihood estimators.

**Citation:** Carvalho L (2013) Bayesian Centroid Estimation for Motif Discovery. PLoS ONE 8(12): e80511. doi:10.1371/journal.pone.0080511

**Editor:** Matteo G. A. Paris, Università degli Studi di Milano (University of Milan), Italy

**Received** June 4, 2013; **Accepted** October 3, 2013; **Published** December 6, 2013



**Funding:** The author is supported by National Science Foundation grant DMS-1107067. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The author has declared that no competing interests exist.

* E-mail: lecarval@math.bu.edu

## Introduction

In motif discovery we are given a set of sequences that share a common motif and aim to identify the motif profile—the frequency of symbols for each position in the pattern—and the positions in each sequence where the motifs are. It is assumed that the motifs have significantly different profiles from sequence background. This problem has gained attention and relevance in the past 25 years mainly due to biological applications; a classical example is regulation of gene expression by transcription factors that bind to specific motifs in genomic promoter regions [1–3]. For this reason, we refer to the positions where the motifs are realized in the sequences as "binding sites".

Due to its importance, hundreds of procedures have been proposed for motif discovery [4,5]. While some approaches seek to characterize motifs and their binding sites using dictionary methods that capture over-representation of words as evidence [6,7], it is common to represent motif compositions by a position weight matrix (PWM) [18] and specify a parametric model where sequences are generated conditionally on motif and background compositions and binding sites. Binding sites can then be regarded as missing data; parameters for the compositions can be estimated using expectation-maximization (EM) [9] in a frequentist setup [10,11], or assigned a prior distribution in a Bayesian setup [12–14]. Other computational approaches are based on evolutionary algorithms and population clustering [15–17].

Even when exploiting prior information in both compositions and binding site configurations in a Bayesian setup, motif discovery is still considered a hard problem since motifs are usually short relative to sequence length and have a composition that might be hard to distinguish from background (see, for instance, [4].) It is then imperative to rely on more refined, informative estimation methods that better glean information from the posterior distribution of binding site configurations. Discrete inferential methods with this goal have recently been proposed, including the median probability model of Barbieri and Berger [8] and the centroid estimator [19,20].

Estimators based on centroid inference, in particular, have been more successful than the ubiquitous maximum *a posteriori* (MAP) estimator when applied to motif discovery in models that account for sequence conservation [21,22]. These estimators, however, were defined from a thresholded loss function that mostly compares binding sites across sequences instead of more finely comparing sequences position-wise for binding site overlaps, as in the traditional centroid estimator (details in Methods.) Moreover, these results rely on sampled binding site configurations to derive the centroid and thus do not offer a characterization of the estimator. Centroid estimators were also shown to yield more compact centered credibility sets then MAP estimators when applied to sequence alignment [23].

In this paper, we propose a novel centroid estimator that arises from a more refined and arguably more natural loss function and that can, in contrast to previous approaches, be fully characterized as a function of marginal posteriors in the space of binding site configurations. In this sense, we argue that the proposed estimator is a better representative of the posterior space of configurations. In addition, as a by-product of its derivation, we obtain informative summaries of the distribution of posterior mass. To this end, we adopt a Bayesian model for motif discovery on multiple sequences with multiple possible binding sites that is an extended version of the classic model from Liu, Neuwald, and Lawrence [14]. The motivation for this extended model is twofold: to obtain a feasible computational method while still retaining a realistic interpretation and to allow us to focus the discussion on the proposed estimator.

## Methods

We present our approach starting from a simple model and building up to the most general setup in the next sections.

## One Sequence, One Binding Site

Suppose we observe a sequence $R$, $|R| \doteq n$, and wish to infer the location of the only binding site $Y$, $Y \in \{1,\ldots,n-L+1\}$. Following the Bayesian model from [14], we assume that there is only one motif of *fixed* length $L$ and that sequences are generated conditionally independently according to a product multinomial model given binding site positions and motif and background compositions. Thus, for an alphabet $\mathcal{S}$, we define $\theta_0 = (\theta_{0,s})_{s\in\mathcal{S}}$ as background probabilities of generating each letter in $\mathcal{S}$ and, for each position $i = 1,\ldots,L$ in the motif, $\theta_i = (\theta_{i,s})_{s\in\mathcal{S}}$ as the probabilities of generating each letter at the $i$-th position in the motif. To simplify the notation we denote $\Theta = (\theta_0,\theta_1,\ldots,\theta_L)$. The likelihood is then:

$$\mathbb{P}(R|\,Y,\Theta) = \prod_{s\in\mathcal{S}}\prod_{j\in BG} \theta_{0,s}^{I(R_j=s)} \prod_{j=1}^{L} \theta_{j,s}^{I(R_{Y-j+1}=s)},$$

where $j \in BG$ means position $j$ in background.

Setting a non-informative prior on $Y$, $\mathbb{P}(Y) = (n-L+1)^{-1}$, we have the posterior:

$$\mathbb{P}(Y|\,R,\Theta) = \frac{\mathbb{P}(R|\,Y,\Theta)\mathbb{P}(Y|\,\Theta)}{\sum_{\tilde{Y}=1}^{n-L+1}\mathbb{P}(R|\,\tilde{Y},\Theta)\mathbb{P}(\tilde{Y}|\,\Theta)} = \frac{\mathbb{P}(R|\,Y,\Theta)}{\sum_{\tilde{Y}=1}^{n-L+1}\mathbb{P}(R|\,\tilde{Y},\Theta)}.$$

One traditional estimator is the maximum *a posteriori* (MAP) estimator,

$$\hat{Y}_M = \operatorname*{argmax}_{\tilde{Y}=1,\ldots,n-L+1} \mathbb{P}(\tilde{Y}|\,R,\Theta),$$

but we argue for an estimator that accounts for differences in positions when comparing binding site configurations. Using Bayesian decision theory [24] we look for an estimator that minimizes, on average, a more refined loss function $H$:

$$\hat{Y}_C = \operatorname*{argmin}_{\tilde{Y}=1,\ldots,n-L+1} \mathbb{E}_{Y|R,\Theta}[H(\tilde{Y},Y)]. \tag{1}$$

We adopt a generalized Hamming loss $H$,

$$H(\tilde{Y},Y) = \sum_{i=1}^{n} h(l_i(\tilde{Y}),l_i(Y)),$$

where $l_i(Y)$ returns the "state" of position $i$: if $i$ is a background position, $l_i(Y) = 0$, otherwise $l_i(Y) = i - Y + 1$, that is, $l_i(Y)$ returns the position in the motif. Loss function $H$ compares configurations position-wise according to $h$, which in turn compares states. If we define $m(i) \doteq I(i > 0)$ to indicate if state $i$ is a motif state then one option for $h$ is $h(i,j) \doteq I(m(i) \neq m(j))$, which yields a loss $H$ that accounts for overlap in binding sites. Such metric is commonly adopted to measure binding site level accuracy, as in the performance coefficients in [4,5,25].

Estimator $\hat{Y}_C$ is a *generalized centroid estimator*; for instance, if $h$ is a common zero-one loss, $h(i,j) = I(i \neq j)$, $H$ corresponds to Hamming loss, and thus $\hat{Y}_C$ is the regular centroid estimator [19,20]. As Carvalho and Lawrence [26] argue, centroid estimators more effectively represent the space since they are closer to posterior means; in contrast, it can be shown that $\hat{Y}_M$ arises from a zero-one loss function which yields the posterior mode [26].

Let us now derive more specific expressions for $H$ and $\hat{Y}_C$. We first notice that if $|\tilde{Y} - Y| \geq L$ then the binding sites do not overlap and so $H(\tilde{Y},Y) = 2\sum_{j=1}^{L} h(j,0) \doteq H^*$, the null overlap distance between two configurations. Alternatively, when $|\tilde{Y} - Y| < L$ then

$$\begin{aligned} H(\tilde{Y},Y) = &\sum_{j=1}^{|\tilde{Y}-Y|} h(j,0) + \sum_{j=L-|\tilde{Y}-Y|+1}^{L} h(j,0) + \\ &\sum_{j=1}^{L-|\tilde{Y}-Y|} h(j,j+|\tilde{Y}-Y|), \end{aligned} \tag{2}$$

since the common backgrounds in $\tilde{Y}$ and $Y$ do not affect $H(\tilde{Y},Y)$, the first two terms above account for the left and right "tails" where binding sites in one sequence are matched with background in the other sequence, and the last term accounts for the overlap in binding sites. We also note that $H(\tilde{Y},Y)$ is actually a function of $|\tilde{Y} - Y|$.

Instead of a loss function we can also define our estimator in terms of a *gain* function $G(\tilde{Y},Y) \doteq 1 - H(\tilde{Y},Y)/H^*$. Note that $0 \leq G(\tilde{Y},Y) \leq 1$; in particular, when $|\tilde{Y} - Y| \geq L$ there is no gain, $G(\tilde{Y},Y) = 0$, and if $\tilde{Y} = Y$ we have $G(\tilde{Y},Y) = 1$. As a consequence, we can simply write $G(\tilde{Y},Y) = I(|\tilde{Y} - Y| < L)(1 - H(\tilde{Y},Y)/H^*)$ with $H$ from Equation 2. Noting that $G$, like $H$, is also a function of $|\tilde{Y} - Y|$, we obtain the following characterization:

**Theorem 1** *The centroid estimator* $\hat{Y}_C$ *is*

$$\hat{Y}_C = \operatorname*{argmax}_{\tilde{Y}=1,\ldots,n-L+1} G(\tilde{Y},\cdot) * \mathbb{P}(\cdot|\,R,\Theta),$$

a *convolution between* $G$ *and the posterior distribution on* $Y$.

*Proof.* The result follows directly from the definition in Equation 1:

$$\begin{aligned} \hat{Y}_C &= \operatorname*{argmin}_{\tilde{Y}=1,\ldots,n-L+1} \mathbb{E}_{Y|R,\Theta}[H(\tilde{Y},Y)] \\ &= \operatorname*{argmax}_{\tilde{Y}=1,\ldots,n-L+1} \mathbb{E}_{Y|R,\Theta}[I(|\tilde{Y} - Y| < L)(1 - H(\tilde{Y},Y)/H^*)] \\ &= \operatorname*{argmax}_{\tilde{Y}=1,\ldots,n-L+1} \sum_{Y=\max\{1,\tilde{Y}-L+1\}}^{\min\{n-L+1,\tilde{Y}+L-1\}} G(\tilde{Y},Y)\mathbb{P}(Y|\,R,\Theta) \\ &= \operatorname*{argmax}_{\tilde{Y}=1,\ldots,n-L+1} G(\tilde{Y},\cdot) * \mathbb{P}(\cdot|\,R,\Theta), \end{aligned}$$

as required.

When contrasted to $\hat{Y}_M$ we can see the effect of having a higher resolution loss function: $\hat{Y}_C$ gathers probability support from nearby, relative to $H$, binding site configurations instead of just picking the most likely configuration. More specifically, for $h$ that corresponds to the overlap in binding sites, we have $H^* = 2L$, $H(\tilde{Y},Y) = 2\min\{|\tilde{Y} - Y|, L\}$, and so $G(\tilde{Y},Y) = I(|\tilde{Y} - Y| < L)(1 - |\tilde{Y} - Y|/L) = \max\{0, 1 - |\tilde{Y} - Y|/L\}$, a "step pyramid" convolution filter that weights farther contributions less heavily. *From now on we will be adopting this loss/gain function.*

Other choices of $G$ could be used, but they do not necessarily correspond to a generalized Hamming loss, and thus not to a centroid estimator (as defined here) either. The centroid estimator in [21,22], for instance, adopts the thresholded gain $G(\tilde{Y},Y) = I(|\tilde{Y} - Y| < L/2)[1 - I(\tilde{Y} \neq Y)\varepsilon]$, close to a "half pla-

teau" filter, where $\varepsilon$ is an infinitesimal meant to break ties. Besides offering less resolution when comparing binding site configurations—binding sites are only compared for a "significant" overlap—this gain function does not result from a generalized Hamming loss since positional information is needed to assess if an overlap is significant or not. Finally, to get some insight into the new estimator, check the first example in the Results section.

## One Sequence, Multiple Binding Sites

We now allow for multiple binding sites by defining $Y=\{Y_k\}$ as the collection of *non-overlapping* binding sites $Y_k$. The likelihood is similar, but accounts for the multiple binding sites:

$$\mathbb{P}(R|\,Y,\Theta)=\prod_{s\in\mathcal{S}}\prod_{i\in BG}\theta_{0,s}^{I(R_i=s)}\prod_{k=1}^{|Y|}\prod_{i=1}^{L}\theta_{i,s}^{I(R_{Y_k+i-1}=s)}.$$

Given the "entropic" effect of possibly having many binding sites, we need to adopt a better prior for $Y$ that takes into account the number of possible configurations for the binding sites. So, instead of naively electing $\mathbb{P}(Y)\propto 1$, we explore a hierarchical structure: if $c(Y)=|Y|$, the number of binding sites in $Y$, we note that $\mathbb{P}(Y)=\mathbb{P}(Y,c(Y))=\mathbb{P}(Y\,|\,c(Y))\mathbb{P}(c(Y))$, then first set $\mathbb{P}(Y\,|\,c(Y))\propto 1$—binding site configurations are equally likely given the number of binding sites—and next define $\mathbb{P}(c(Y))$.

In what follows we settle on a prior distribution for $c(Y)$ that is based on a Markov chain with two states, background and motif, where the probability of transitioning to background, either from background or motif, and of starting at background is $p$; this approach results in

$$\mathbb{P}(c(Y))\propto\binom{n-c(Y)(L-1)}{c(Y)}p^{n-c(Y)L}(1-p)^{c(Y)},\qquad(3)$$

since there needs to be $c(Y)$ transitions to the motif state. This prior structure offers a good degree of flexibility through $p$: we can further set a hyperprior distribution on $p$, or specify it directly based on the expected number $b$ of binding sites in the sequence; if $n$ is large compared to $b$, as usual, then $p$ should be close to one, $c(Y)$ is approximately Poisson with mean $n(1-p)$ and thus $p\doteq 1-b/n$ becomes a good candidate.

Going back to our inferential goal we note that, in contrast to the one binding site case from last section, posterior inference is more difficult since comparing configurations with different number of binding sites is not amenable to a systematic approach. Our first approximation is to consider local estimators for each group of configurations with a fixed number of binding sites and then appeal to a triangle inequality:

$$H(Y,\hat{Y})\le H(Y,\hat{Y}_c)+H(\hat{Y}_c,\hat{Y}),$$

where $Y$ is a configuration with $c$ binding sites, $\hat{Y}_c$ is the constrained estimator for all configurations with $c$ binding sites, and $\hat{Y}$ is the (overall) centroid estimator. Let $C\doteq\lfloor n/L\rfloor$ be the maximum number of binding sites in $R$, and recall that for the centroid estimator we wish to find $\tilde{Y}$ that minimizes

$$\mathbb{E}_{Y|R,\Theta}[H(\tilde{Y},Y)]=\sum_{c=0}^{C}\sum_{Y:c(Y)=c}H(\tilde{Y},Y)\mathbb{P}(Y\,|\,R,\Theta).$$

Using the triangle inequality for each group we then have

$$\begin{aligned}\mathbb{E}_{Y|R,\Theta}[H(\tilde{Y},Y)]&\le\sum_{c=0}^{C}\sum_{Y:c(Y)=c}[H(\tilde{Y},\tilde{Y}_c)+H(\tilde{Y}_c,Y)]\mathbb{P}(Y\,|\,R,\Theta)\\&=\sum_{c=0}^{C}[H(\tilde{Y},\tilde{Y}_c)+\sum_{Y:c(Y)=c}H(\tilde{Y}_c,Y)\mathbb{P}(Y|c(Y)\\&\quad=c,R,\Theta)]\mathbb{P}(c(Y)=c|R,\Theta),\end{aligned}\qquad(4)$$

where $\tilde{Y}_c$ is an arbitrary point in $\{Y:c(Y)=c\}$. Our task is now to find an estimator—let us still call it centroid—that minimizes the right-hand bound in Equation 4 above. This goal suggests a two-step strategy:

1. For each number of binding sites, $c=1,\ldots,C$, find the *local* centroids

$$\hat{Y}_c=\underset{\tilde{Y}:c()=c}{\arg\min}\,\mathbb{E}_{Y\,|\,c(Y)=c,R,\Theta}[H(\tilde{Y},Y)]\qquad(5)$$

as the $\tilde{Y}_c$ in Equation 4.

2. Find the *global* centroid given the local centroids $\{\hat{Y}_c\}_{c=1}^{C}$,

$$\hat{Y}=\underset{\tilde{Y}}{\arg\min}\,\mathbb{E}_{c(Y)\,|\,R,\Theta}[H(\hat{Y}_{c(Y)},\tilde{Y})].\qquad(6)$$

We note that this strategy does not guarantee that the bound is minimized; the main goal here is computational convenience. Let us tackle each step of this heuristic next. To this end we need $\mathbb{P}(c(Y)|\,R,\Theta)$ and marginal posteriors $\mathbb{P}(Y_k|\,c(Y),R,\Theta)$; obtaining these posterior probabilities is a standard procedure, but we provide details on how they can be computed in File S1 for completeness.

## Local Centroids

Even when the number of binding sites is fixed, minimizing the conditional posterior expectation of $H(\tilde{Y},Y)$ can be challenging: we would still have to consider for each candidate configuration $\tilde{Y}$ the posterior probability of configurations with all binding sites to the left of the first binding site in $\tilde{Y}$, in-between binding sites in $\tilde{Y}$, and so on. We adopt another approximation and decide to minimize a *paired* Hamming loss $H_A$ where binding site positions are matched according to their order:

$$H_A(\tilde{Y},Y)=\sum_{k=1}^{c(Y)}H_1(\tilde{Y}_k,Y_k),$$

where $H_1(\tilde{Y}_k,Y_k)$ is Hamming loss when comparing sequences with only one binding site at $\tilde{Y}_k$ and $Y_k$, respectively, that is, $H_1(\tilde{Y}_k,Y_k)=2\max\{|\tilde{Y}_k-Y_k|,L\}$. From the definition we have that $H_A$ upper bounds $H$: $H_A(\tilde{Y},Y)\ge H(\tilde{Y},Y)$. As a bad approximation example, if $\tilde{Y}_k=Y_{k+1}$ for $k=1,\ldots,c(Y)-1$ then $H_A(\tilde{Y},Y)=c(Y)L$, since each pair of binding sites $\tilde{Y}_k$ and $Y_k$ does not overlap, while $H(\tilde{Y},Y)=2L$ since only $Y_1$ and $\tilde{Y}_{c(Y)}$ are in disagreement with background.

The next result adapts Theorem 1 to yield the paired local centroids.

**Lemma 2** *If $\mathbb{P}_k(\cdot\,|\,c(Y)=c,R,\Theta)$ is the marginal conditional posterior on $Y_k$ then the paired local centroids are*

$$\hat{\tilde{Y}}_c = \underset{\tilde{Y}:c(\tilde{Y})=c}{\operatorname{argmax}} \sum_{k=1}^{c} G(\tilde{Y}_k,\cdot) * \mathbb{P}_k(\cdot \mid c(Y)=c,R,\Theta)$$

*Proof.* In the same spirit of Theorem 1, we use the conditional estimator in Equation 5 with the paired loss $H_A$:

$$\begin{aligned}
\hat{\tilde{Y}}_c &= \underset{\tilde{Y}:c(\tilde{Y})=c}{\operatorname{argmin}} \mathbb{E}_{Y \mid c(Y)=c,R,\Theta}[H_A(\tilde{Y},Y)] \\
&= \underset{\tilde{Y}:c(\tilde{Y})=c}{\operatorname{argmin}} \sum_{Y:c(Y)=c} \sum_{k=1}^{c} H_1(\tilde{Y}_k,Y_k)\mathbb{P}(Y \mid c(Y)=c,R,\Theta) \\
&= \underset{\tilde{Y}:c(\tilde{Y})=c}{\operatorname{argmin}} \sum_{k=1}^{c} \sum_{Y_k=(k-1)L+1}^{n-(c-k+1)L+1} H_1(\tilde{Y}_k,Y_k)\mathbb{P}(Y_k \mid c(Y)=c,R,\Theta) \\
&= \underset{\tilde{Y}:c(\tilde{Y})=c}{\operatorname{argmax}} \sum_{k=1}^{c} \sum_{Y_k=\max\{(k-1)L+1,\tilde{Y}_k-L\}}^{\min\{n-(c-k+1)L+1,\tilde{Y}_k+L\}} \\
&\quad G(\tilde{Y}_k,Y_k)\mathbb{P}(Y_k \mid c(Y)=c,R,\Theta) \\
&= \underset{\tilde{Y}:c(\tilde{Y})=c}{\operatorname{argmax}} \sum_{k=1}^{c} G(\tilde{Y}_k,\cdot) * \mathbb{P}_k(\cdot \mid c(Y)=c,R,\Theta),
\end{aligned}$$

and the result follows.

We can spot in Lemma 2 the familiar convolutions, but now with the marginal posteriors $\mathbb{P}(Y_k \mid c(Y),R,\Theta)$ and in a more restricted range. We have a nice characterization, but we still have to optimize a sum to obtain the local centroids; to this end we explore the same recursive structure that allows us to compute forward and backward sums (see File S1 for details.) Let us define $f(\tilde{Y}_k) \doteq G(\tilde{Y}_k,\cdot) * \mathbb{P}_k(\cdot \mid c(Y)=c,R,\Theta)$ as the convolution against the marginal posterior on $Y_k$; then we should have

$$\begin{aligned}
&\max_{\tilde{Y}:c(\tilde{Y})=c} \sum_{k=1}^{c} f(\tilde{Y}_k) = \\
&\max_{\tilde{Y}_c=(c-1)L+1,\ldots,n-cL+1} [f(\tilde{Y}_c) + \max_{\tilde{Y}_1,\ldots,\tilde{Y}_{c-1}} \sum_{k=1}^{c-1} f(\tilde{Y}_k)].
\end{aligned} \tag{7}$$

This important observation allows us to obtain $\hat{\tilde{Y}}_c$ using the dynamic programming approach listed in Algorithm 1, as Theorem 3 formalizes.

**Algorithm 1** Find $\hat{\tilde{Y}}_c$ using dynamic programming.

*Construct partial maxima and backtrack pointers:*

Step 1. Set $m_1(\tilde{Y}_1)=f(\tilde{Y}_1)$ for $\tilde{Y}_1=1,\ldots,n-cL+1$.

Step 2. For $k=2,\ldots,c$ and $\tilde{Y}_k=(k-1)L+1,\ldots,n-(c-k+1)L+1$ do: set backtrack pointers

$$A_{k-1}(\tilde{Y}_k) = \underset{\tilde{Y}_{k-1}=(k-2)L+1,\ldots,\tilde{Y}_k-L}{\operatorname{argmax}} m_{k-1}(\tilde{Y}_{k-1}).$$

and set partial maximum sum $m_k$ as

$$m_k(\tilde{Y}_k)=f(\tilde{Y}_k)+m_{k-1}(A_{k-1}(\tilde{Y}_k)).$$

*Reconstruct centroid $\hat{\tilde{Y}}_c$ using backtrack pointers:*

Step 3. Set last binding site position:

$$\hat{\tilde{Y}}_{c,c} = \underset{\tilde{Y}_c=(c-1)L+1,\ldots,n-L+1}{\operatorname{argmax}} m_c(\tilde{Y}_c).$$

Note that, by construction, $\max_{\tilde{Y}:c(\tilde{Y})=c} \sum_{k=1}^{c} f(\tilde{Y}_k) = m_c(\hat{\tilde{Y}}_{c,c})$.

Step 4. For $k=c,\ldots,2$ do: recover the remainder of $\hat{\tilde{Y}}_c$ by setting $\hat{\tilde{Y}}_{c,k-1}=A_{k-1}(\hat{\tilde{Y}}_{c,k})$.

**Theorem 3** *Algorithm 1 correctly identifies the paired local centroids*

$$\hat{\tilde{Y}}_c = \underset{\tilde{Y}:c(\tilde{Y})=c}{\operatorname{argmin}} \mathbb{E}_{Y \mid c(Y)=c,R,\Theta}[H_A(\tilde{Y},Y)].$$

*Proof.* From Lemma 2 we know that $\hat{\tilde{Y}}_c$ is the argument of $\max_{\tilde{Y}:c(\tilde{Y})=c} \sum_{k=1}^{c} f(\tilde{Y}_k)$. The key device in Algorithm 1 is to exploit the recursion in Equation 7 to define $m_1(\tilde{Y}_1)=f(\tilde{Y}_1)$ and

$$m_k(\tilde{Y}_k)=f(\tilde{Y}_k)+\max_{\tilde{Y}_{k-1}=(k-2)L+1,\ldots,\tilde{Y}_k-L} m_{k-1}(\tilde{Y}_{k-1}), \tag{8}$$

for $k>1$, to store partial sum maxima. Now it follows that

$$\max_{\tilde{Y}:c(\tilde{Y})=c} \sum_{k=1}^{c} f(\tilde{Y}_k) = \max_{\tilde{Y}_c=(c-1)L+1,\ldots,n-cL+1} m_c(\tilde{Y}_c),$$

and so Step 3 must be correct. The correctness of Step 4 relies on the right specification of $m$ in Steps 1 and 2; but these steps are a straightforward application of Equation 7 using the definition of $m_1$ and a formulation of Equation 8 based on the backtrack pointers $A$, and so the algorithm is correct.

We note that the paired local centroids minimize an expected posterior upper bound $H_A$ on the loss $H$, and so the actual local centroid might not be attained. We expect, however, that for common cases in which the motif coverage $c(Y)L$ is much smaller than $n$ that the bound is tight since $H_A$ approximates $H$ well and thus the two local centroids often coincide.

## Global Centroid

While the local centroids already convey information about the distribution of posterior mass in the space of binding site configurations, the end goal of the analysis is a point estimate that is, in itself, a good representative of the space. Following the strategy we outlined in the beginning of this section, we can further summarize the information in the local centroids by identifying a configuration $\hat{\tilde{Y}}$ that minimizes the expected conditional Hamming loss, as in Equation 6. This approach, however, entails the same difficulties as defining the centroid based on all points in the space, and it is thus not treatable by a systematic approach—we are now just restricting the configurations to the local centroids.

The global centroid can be defined by direct enumeration of all possible configurations while keeping the minimizer of the expected conditional posterior loss, but this "brute-force" approach considers an exponential number of solutions. A simple heuristic is to restrict the global centroid to be one of the local centroids,

$$\hat{\tilde{Y}} = \underset{\tilde{Y} \in \{\hat{\tilde{Y}}_c\}_{c=0}^{C}}{\operatorname{argmin}} \mathbb{E}_{c(Y) \mid R,\Theta}[H(\hat{\tilde{Y}}_{c(Y)},\tilde{Y})]. \tag{9}$$

Another alternative is to just take as global centroid the local

centroid of the modal number of binding sites, $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{c^*}$, where $c^* \doteq \underset{c=0,\ldots,C}{\arg\max}\, \mathbb{P}(c(Y)=c \mid R,\Theta)$. From now on we adopt the global centroid in Equation 9 for simplicity and, again, computational expediency.

## Constrained Global Centroid

A *constrained*, on the number of binding sites, global centroid might be more computationally feasible since we are restricting the space of available configurations. For instance, consider the 1-global centroid,

$$\hat{\boldsymbol{Y}}_o \doteq \underset{\tilde{\boldsymbol{Y}}:c(\tilde{\boldsymbol{Y}})=1}{\arg\min}\, \mathbb{E}_{Y \mid R,\Theta}[H(\tilde{\boldsymbol{Y}},Y)].$$

As when defining local centroids, we can approximate $\hat{\boldsymbol{Y}}_o$ using a paired loss, and since

$$\begin{aligned}
\mathbb{E}_{Y \mid R,\Theta}[H_A(\tilde{\boldsymbol{Y}},Y)] &= \sum_{c=0}^{C} \sum_{Y:c(Y)=c} \sum_{k=1}^{c} H_1(\tilde{\boldsymbol{Y}},Y_k)\mathbb{P}(Y \mid R,\Theta) \\
&= \sum_{i=1}^{n} \sum_{c=0}^{C} \sum_{Y:c(Y)=c} \sum_{k=1}^{c} H_1(\tilde{\boldsymbol{Y}},i)\mathbb{P}(Y_k=i \mid R,\Theta) \\
&= \sum_{i=1}^{n} H_1(\tilde{\boldsymbol{Y}},i) \sum_{c=0}^{C} \sum_{Y:c(Y)=c} \sum_{k=1}^{c} \mathbb{P}(Y_k=i \mid R,\Theta) \\
&= \sum_{i=1}^{n} H_1(\tilde{\boldsymbol{Y}},i) P_c(i \mid R,\Theta),
\end{aligned}$$

where

$$P_c(i \mid R,\Theta) \doteq \sum_{c=1}^{C} \sum_{Y:c(Y)=c} \sum_{k=1}^{c} \mathbb{P}(Y_k=i \mid R,\Theta), \qquad (10)$$

we have that

$$\hat{\boldsymbol{Y}}_o = \underset{\tilde{\boldsymbol{Y}}:c(\tilde{\boldsymbol{Y}})=1}{\arg\min}\, \mathbb{E}_{Y \mid R,\Theta}[H_A(\tilde{\boldsymbol{Y}},Y)] = \underset{\tilde{\boldsymbol{Y}}:c(\tilde{\boldsymbol{Y}})=1}{\arg\max}\, G(\tilde{\boldsymbol{Y}},\cdot) * P_c(\cdot \mid R,\Theta).$$

It is important to note that while the restriction of one binding site might seem artificial, the derivation of $\hat{\boldsymbol{Y}}_o$ is helpful in recognizing sequence regions that are likely to host binding sites. In fact, since $P_c$ captures the posterior probability of having a binding site starting at each position, and considering the overlap gain $G$, the convolution of $G$ and $P_c$ highlights positions that have higher posterior probability of being covered by a binding site.

## Multiple Sequences, Multiple Binding Sites per Sequence, Random Motif

We are now ready to address our model in broader generality: the dataset now comprises $m$ sequences, $R=\{R_i\}_{i=1}^m$, and thus binding site configurations are also indexed by sequence, $Y=\{Y_i\}_{i=1}^m$. As before, we have that $Y$ is independent of motif parameters $\Theta$, but we further assume that sequences and configurations are conditionally independent given $\Theta$:

$$\mathbb{P}(R,Y \mid \Theta) = \prod_{i=1}^{m} \mathbb{P}(R_i,Y_i \mid \Theta) = \prod_{i=1}^{m} \mathbb{P}(R_i \mid Y_i,\Theta)\mathbb{P}(Y_i). \qquad (11)$$

Given $\Theta$ we would be able to apply the methods discussed this far to each sequence separately: compute forward and backward sums to obtain marginal posterior probabilities for each $Y_i$ and then find local centroids and the $i$-th global centroid. We will, however, assume that $\Theta$ is random, say,

$$\theta_j \sim \mathrm{Dir}(\alpha_j), \quad j=0,1,\ldots,L, \qquad (12)$$

independently with the usual conjugacy [14], and we thus wish to also conduct inference on background and motif compositions. This assumption, albeit more realistic, complicates matters, since the marginal unconditioned posterior distributions of $Y$ and $\Theta$ are not readily available; we are now required to estimate them before obtaining centroid estimates.

To obtain the centroids we follow the procedure described in the last section, but adopting Monte Carlo estimates of the marginal posterior distributions, for $i=1,\ldots,m$,

$$\begin{aligned}
\widehat{\mathbb{P}}(c(Y_i)=c \mid R) &\approx \frac{1}{T}\sum_{t=1}^{T} I(c(Y_i^{(t)})=c), \\
\widehat{\mathbb{P}}(Y_{ik}=j \mid c(Y_i)=c,R) &\approx \frac{\sum_{t=1}^{T} I(Y_{ik}^{(t)}=j)I(c(Y_i^{(t)})=c)}{\sum_{t=1}^{T} I(c(Y_i^{(t)})=c)}, \\
& \qquad k=1,\ldots,c,
\end{aligned}$$

where $T$ is the number of samples. In File S1 we present a Gibbs sampler [27,28] that draws $Y_i$ for each sequence given $\Theta$ and then samples $\Theta$ conditional on the binding site configurations $Y$, similar to the approach in [14].

# Results

## Illustrative Examples

**Example 1: One Sequence, One Binding Site.** Consider the following sequence of length $n=200$ from the nucleotide alphabet $\mathcal{S}=\{\mathrm{A, C, G, T}\}$,

```
      10   20   30   40   50
      |    |    |    |    |
GCCACTTTCGGGCCCGTGTCTAACGCACCACGGGC-
TACGTGACGGTGTGG CTCTATACTGACGACGTGAAC-
CAAGCTTTACTGAAGGACTTGCTGTTCCC CGACCCA-
TTTCCTGCCAGAACCTCTGACCAGTGTCTAGGGCTAT-
CGCCCG TGATGTCTCATGGCGACGCGCGAGGCGGT-
TGCTCGCCTCACTCCGTTCTG
```

and a motif of length $L=6$ with parameters $\Theta$ given by Table 1. Figure 1 shows the conditional marginal posterior $\mathbb{P}(Y \mid R,\Theta)$ and the convolution $G * \mathbb{P}(\cdot \mid R,\Theta)$ used to obtain the centroid $\hat{\boldsymbol{Y}}_C=36$, binding at the subsequence TACGTG, close to the consensual motif. Note that since $\Theta$ is very informative the posterior profile has clear peaks and in this case $\hat{\boldsymbol{Y}}_c=\hat{\boldsymbol{Y}}_M$, the two estimators coincide.

**Example 2: One Sequence, Multiple Binding Sites.** We revisit the same sequence from Example 1, but now allow for at most $C=\lfloor n/L \rfloor=33$ binding sites, and adopt the prior given in Equation 3 with $b=3$ and thus $p=1-b/n=0.985$. Using Algorithm 1 in File S1 we are able to compute the conditional marginal posteriors $\mathbb{P}(c(Y) \mid R,\Theta)$ and $\mathbb{P}(Y_k \mid c(Y),R,\Theta)$ for

**Table 1.** Background and motif compositions.

| $\mathcal{S}$ | $\theta_0$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ | $\theta_6$ |
|---|---|---|---|---|---|---|---|
| A | 0.2 | 0.1 | 0.7 | 0.1 | 0.1 | 0.1 | 0.1 |
| C | 0.3 | 0.7 | 0.1 | 0.7 | 0.1 | 0.1 | 0.1 |
| G | 0.3 | 0.1 | 0.1 | 0.1 | 0.7 | 0.1 | 0.7 |
| T | 0.2 | 0.1 | 0.1 | 0.1 | 0.1 | 0.7 | 0.1 |

Background is assumed to be CG-rich, while the motif represents a canonical palindromic E-box, CACGTG [34].
doi:10.1371/journal.pone.0080511.t001

$k=1,\ldots,c(Y)$. These posterior distributions yield the local centroids—according to Algorithm 1—and the global centroid from Equation 9. In Table 2 we list the marginal posterior $\mathbb{P}(c(Y)=c\,|\,R,\Theta)$ up to the smallest $c$ such that $\mathbb{P}(c(Y)\leq c\,|\,R,\Theta)>0.95$, along with the local centroids; the global centroid $\hat{Y}_C$ is highlighted. Interestingly, the global centroid coincides with the local centroid from the modal number of binding sites.

In Figure 2 we display the posterior probabilities of binding site coverage $P_c$ from Equation 10, along with the convolutions that are needed to define the 1-global centroid $\hat{Y}_o=36$. As can be seen, position 36 has a lot of support, being present in all the local centroids listed in Table 2; in fact, the probability of a binding site starting at position 36 is greater than 50%.

While $P_c$ can provide us guidance for which positions are likely to start a binding site, using $P_c$ to define local centroids can be misleading. For instance, we could expect that the local centroid with three binding sites—the modal number of binding sites—would be, following a decreasing order on $P_c$, 36, 63, and 147. However, if we examine the marginal posteriors $\mathbb{P}(Y_k\,|\,c(Y)=3,R,\Theta)$ in Figure 3 we realize that position 13 is

**Table 2.** Centroids and marginal posterior distribution of number of binding sites.

| $c$ | $\hat{Y}_c$ | $\mathbb{P}(c(Y)=c\,\|\,R,\Theta)$ | $\mathbb{P}(c(Y)\leq c\,\|\,R,\Theta)$ |
|---|---|---|---|
| 0 | – | 0.014 | 0.014 |
| 1 | 36 | 0.075 | 0.089 |
| 2 | 36,147 | 0.181 | 0.270 |
| **3** | **13,36,147** | **0.254** | 0.524 |
| 4 | 13,36,63,147 | 0.233 | 0.757 |
| 5 | 13,36,63,147,167 | 0.147 | 0.904 |
| 6 | 3,29,36,63,147,167 | 0.067 | 0.971 |

The global centroid and the modal number of binding sites are highlighted in bold.
doi:10.1371/journal.pone.0080511.t002

favored over position 63 because, if $F_k \doteq G * \mathbb{P}_k(\cdot\,|\,c(Y)=3,R,\Theta)$, $F_1(13)+F_2(36)>F_1(36)+F_2(63)$.

**Example 3: Multiple Sequences, Multiple Binding Sites per Sequence.** For the random motif version of the last example we simulate $m=20$ sequences of same length $n=200$ using $\Theta$ from Table 1 and the prior for $Y_i$, $i=1,\ldots,m$, from Equation 3 with $p=1-1/n=0.995$. We continue focusing on the inference of binding site configurations in the same sequence from previous examples, which is the first sequence in the simulated dataset. We assume a non-informative prior on $\Theta$ by setting $\alpha_{j,s}=1$ for $s\in\mathcal{S}$ and $j=0,\ldots,L$; the prior on each sequence $Y_i$ is the same prior from Example 2 with $p=0.985$. Algorithm 2 in File S1 is run for 10,000 iterations to guarantee convergence (diagnostics not shown.)

The marginal posterior distribution of $\Theta$ can be assessed in Figure 4. Since most positions in the sequences are background sequences $\theta_0$ has very small posterior variances. Also note that the canonical palindromic E-box motif, with consensus CACGTG, is

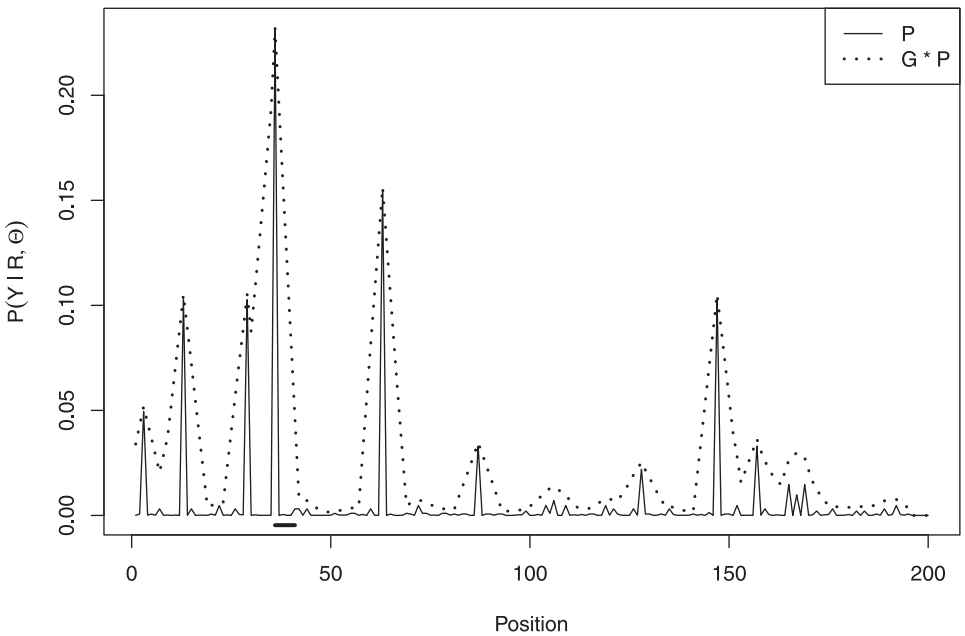


**Figure 1. Conditional marginal probability distribution $\mathbb{P}(Y\,|\,R,\Theta)$ in solid line and convolution $G*\mathbb{P}(\cdot\,|\,R,\Theta)$ in dotted line.** The black thick line close to the axis marks the binding site corresponding to the centroid $\hat{Y}_C$.
doi:10.1371/journal.pone.0080511.g001

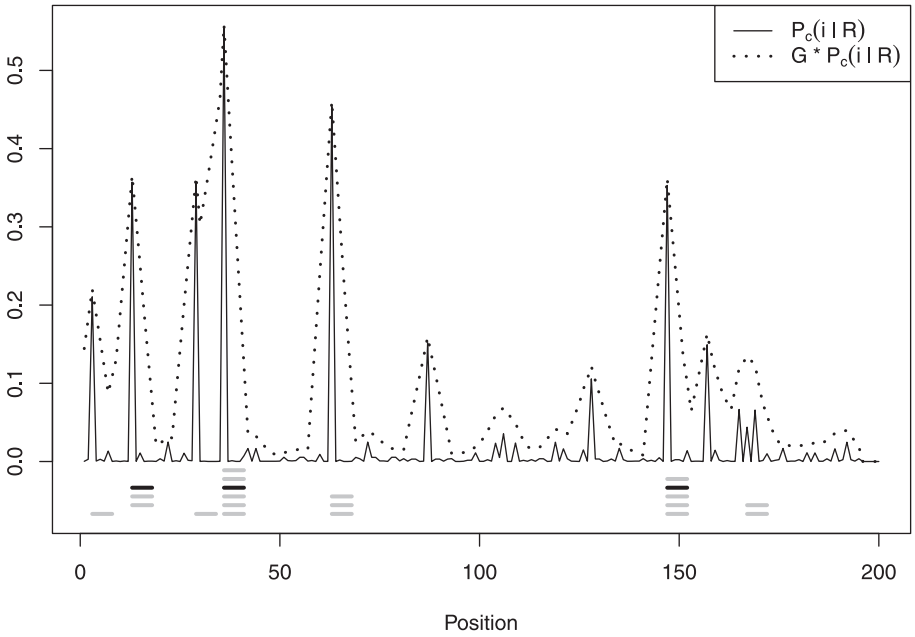


**Figure 2. Posterior binding site coverage $P_c$ in solid line and convolution $G * P_c$ in dotted line.** Local centroids are listed below in gray; the global centroid is in black.
doi:10.1371/journal.pone.0080511.g002

recovered. The procedure is now similar to what we presented in Example 2; the main difference is that the marginal posterior distributions are estimated from Markov chain Monte Carlo (MCMC) samples obtained as shown in File S1. Table 3 lists the estimated marginal posterior distribution of the number of binding sites, the local and global centroids. The global centroid does not coincide with the local centroid for the modal number of binding sites. Moreover, the local centroids here are different from the (conditional) local centroids in the last example, most likely due to the randomness of $\Theta$ being taken into account.

Figure 5 displays the estimated $P_c$, $G * P_c$, and the centroids. We see that compared to the second example some posterior mass has shifted to positions 29 and to the group of positions 166, 167, and 168. Here we clearly see the advantage of a centroid estimator: $G * P_c$, and later $G * \mathbb{P}_k(\cdot \mid R)$, gathers evidence of motif binding from nearby positions, yielding a better summary—

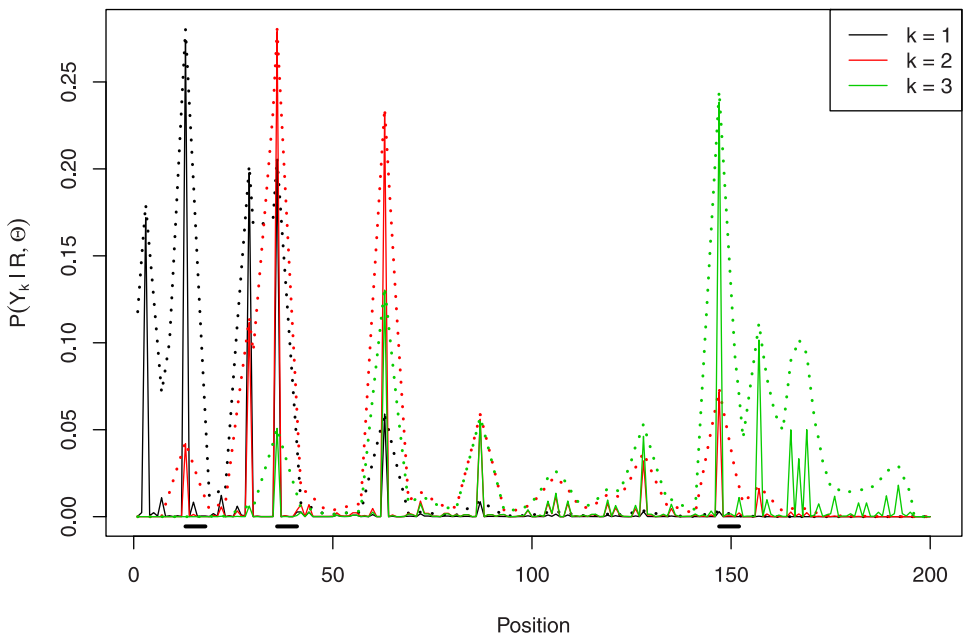


**Figure 3. Marginal posterior distributions $\mathbb{P}(Y_k \mid c(Y)=3,R,\Theta)$ in solid line and convolutions $G * \mathbb{P}(\cdot \mid c(Y),R,\Theta)$ in dotted line.** The local centroid is displayed at the bottom.
doi:10.1371/journal.pone.0080511.g003

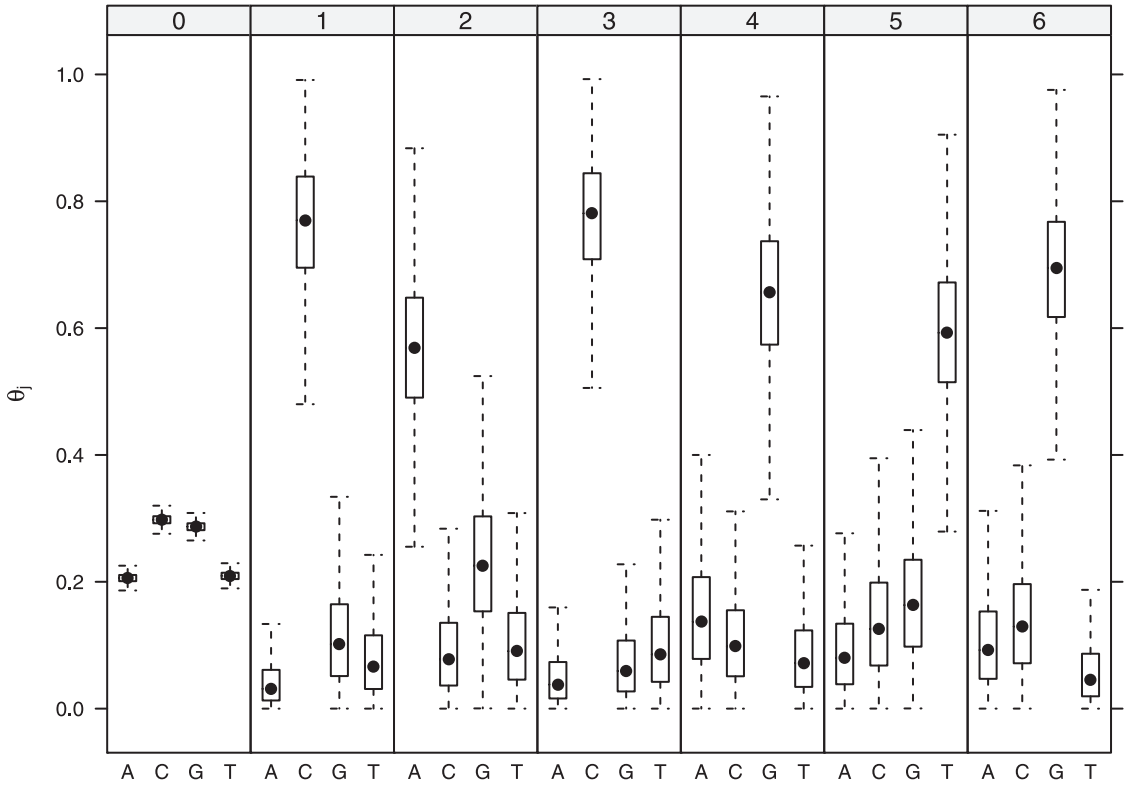


**Figure 4. Boxplots of MCMC samples for $\Theta$.** (Outliers are not shown).
doi:10.1371/journal.pone.0080511.g004

according to our choice of loss function—of the distribution of posterior mass.

The selection of position 167 in the second local centroid $\hat{Y}_2$ might seem puzzling since the peaks at positions 36, 63, and 147 hold higher coverage probabilities. Checking $\hat{\mathbb{P}}(Y_k|R)$ in Figure 6 helps dismiss any doubts: most of the support for these positions come from configurations with higher number of binding sites, as evidenced by the respective local centroids, but these configurations hold relatively low posterior mass. When $c(Y)=2$, the prior on $Y_{2,2}$ assigns more posterior probability to higher positions, close to the end of the sequence, simply because there are more configurations for $Y_{2,2}$ on these positions. It is also important to notice that while none of the positions in the cluster 166–168 has higher marginal posterior mass than positions 63 and 147, the convolution $G * \hat{\mathbb{P}}_2(\cdot|R)$ is maximized at position 167, that is, the cluster when taken together has more support from the data, as weighted by $G$.

**Table 3.** Centroids and estimated marginal posterior distribution of number of binding sites.

| $c$ | $\hat{Y}_c$ | $\hat{\mathbb{P}}(c(Y)=c\|R,\Theta)$ | $\hat{\mathbb{P}}(c(Y)\leq c\|R,\Theta)$ |
|---|---|---|---|
| 0 | – | 0.026 | 0.026 |
| 1 | 29 | 0.107 | 0.133 |
| **2** | **29,167** | 0.210 | 0.343 |
| 3 | 29,63,167 | **0.274** | 0.617 |
| 4 | 13,36,147,167 | 0.201 | 0.818 |
| 5 | 13,29,63,147,167 | 0.120 | 0.938 |
| 6 | 13,29,36,63,147,167 | 0.046 | 0.984 |

The global centroid and the modal number of binding sites are highlighted in bold.
doi:10.1371/journal.pone.0080511.t003

## Case Study

We end this section with an example from the real-world dataset in [5], sequence set yst02r. The dataset contains $m=4$ sequences each with $n=500$ letters. We set $L=16$ and adopt a non-informative prior on $\Theta$, as in the previous example, and the prior on each $Y_i$, for the $i$-th sequence, from Equation 3 with $b=3$ per thousand positions, so $p=1-3/1000=0.997$. As in the previous example, 10,000 iterations suffice to reach convergence.

Let us focus on the second sequence. Figure 7 pictures the binding site coverage probabilities, along with the local centroids. The global centroid $\hat{Y}_C=\{85,105,169\}$ contains three binding sites, and it is also the local centroid for the modal number of binding sites, with $\hat{\mathbb{P}}(c(Y)=3|R)=0.32$. Since most of the posterior mass in concentrated in configurations with $c(Y)=3$, the posterior profiles $\hat{\mathbb{P}}(Y_k|c(Y)=3,R)$ are similar to $P_c$ and are thus omitted.

From the MCMC samples we can produce the MAP estimate $\hat{Y}_M=\{86,105,174\}$ as the configuration with highest frequency among the samples: $\hat{\mathbb{P}}(\hat{Y}_M|R)=0.032$. In fact, we can estimate the posterior probability of each sampled binding site configuration and then, using classic multidimensional scaling [29], visualize the estimated posterior distribution in Figure 8. It is interesting to note that the null configuration—that is, without binding sites—is also very likely with posterior probability 0.024. In contrast, the global centroid has very small posterior probability, close to 0.001;

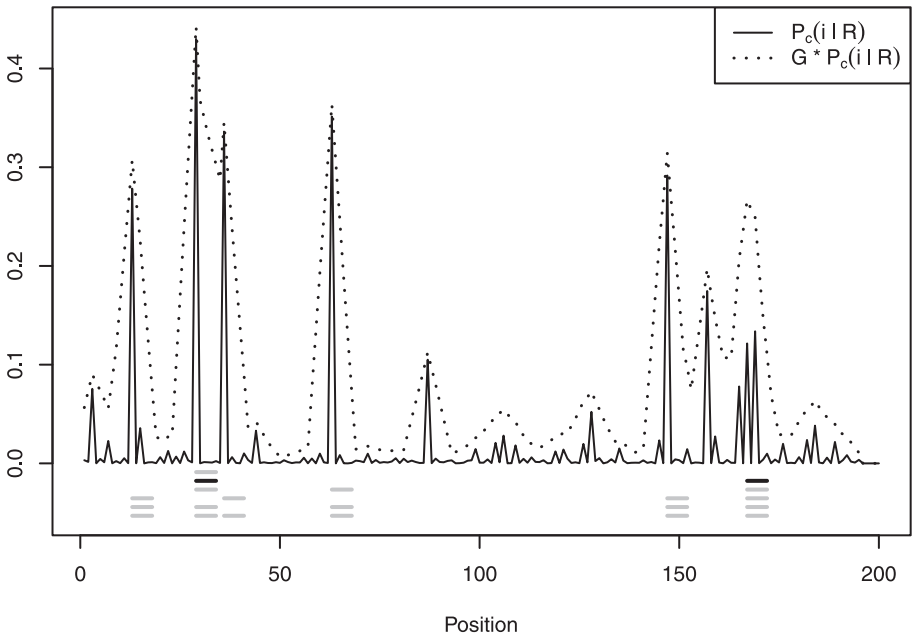


**Figure 5. Estimated posterior binding site coverage $P_c$ in solid line and convolution $G * P_c$ in dotted line.** Local centroids are listed below in gray; the global centroid is in black.
doi:10.1371/journal.pone.0080511.g005

it sits, however, closer to configurations with high posterior mass, including the local centroids with one, two, and four binding sites.

To better assess how the centroid estimator is closer to a mean than a mode estimator, we plot the estimated posterior distribution of the generalized loss function $H$ centered at both $\hat{Y}_C$ and $\hat{Y}_M$ in Figure 9. Since $\mathbb{E}_{Y|R}[H(\hat{Y}_M, Y)] = 42.40$ and $\mathbb{E}_{Y|R}[H(\hat{Y}_C, Y)] = 40.22$, we see that the binding sites in the centroid configuration are, on average, overlapping two extra positions with the binding sites in all the configurations when compared to the MAP estimate's binding sites. Both estimates are fairly similar, but the centroid reminds us that placing the third binding site at position 169, instead of 174, yields an unlikely configuration, but with a higher chance of overlapping with binding sites in positions 160–175 that have high posterior probability. In the context of Figures 8 and 9, the centroid places itself between two clusters that

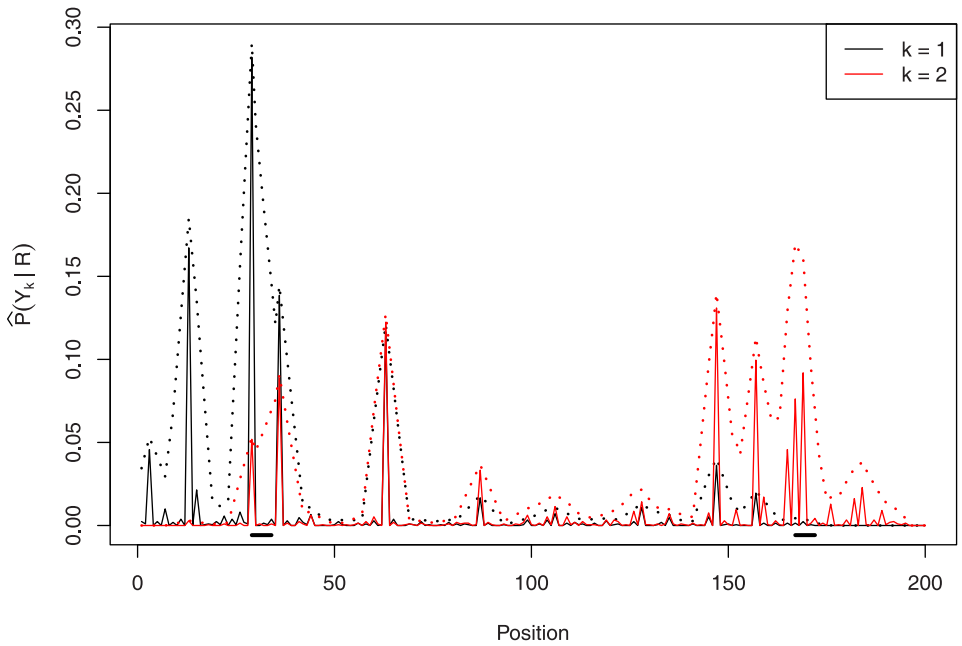


**Figure 6. Estimated marginal posterior distributions $\mathbb{P}(Y_k | c(Y) = 2, R, \Theta)$ in solid line and convolutions $G * \mathbb{P}(\cdot | c(Y), R, \Theta)$ in dotted line.** The local centroid is displayed at the bottom.
doi:10.1371/journal.pone.0080511.g006

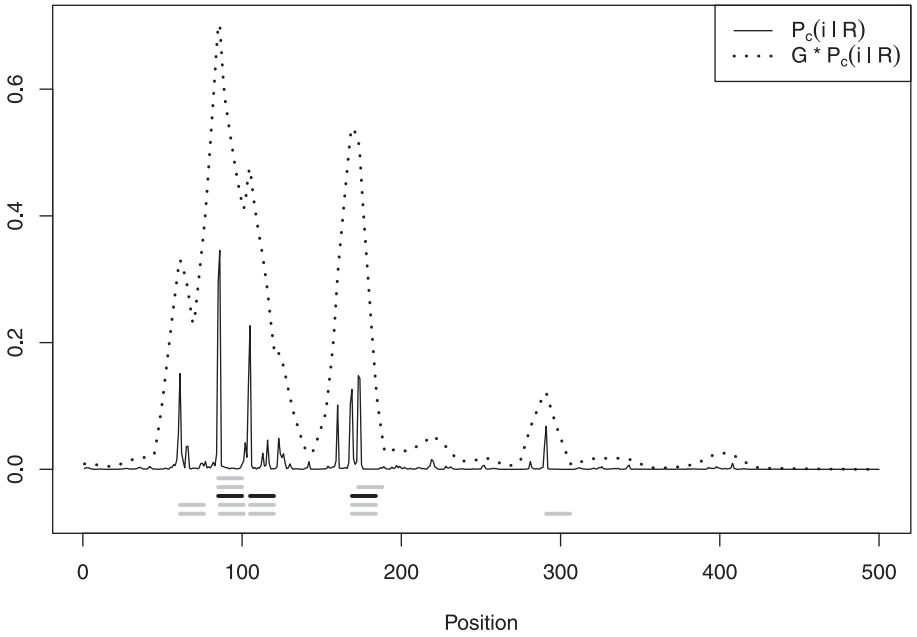


**Figure 7. Estimated posterior binding site coverage $P_c$ and convolution $G * P_c$ for real-world dataset, second sequence.** Binding site coverage $P_c$ in solid line and convolution $G * P_c$ in dotted line. Local centroids are listed below in gray; the global centroid is in black.
doi:10.1371/journal.pone.0080511.g007

concentrate posterior mass: one with configurations $Y$ such that $25 \le H(\hat{Y}_C, Y) \le 40$ and another with configurations further away, satisfying $40 \le H(\hat{Y}_C, Y) \le 50$.

## Discussion

In this paper we have presented a Bayesian approach, similar to the Gibbs motif sampler in [12,14], that jointly models motif and background compositions and binding site locations in a set of

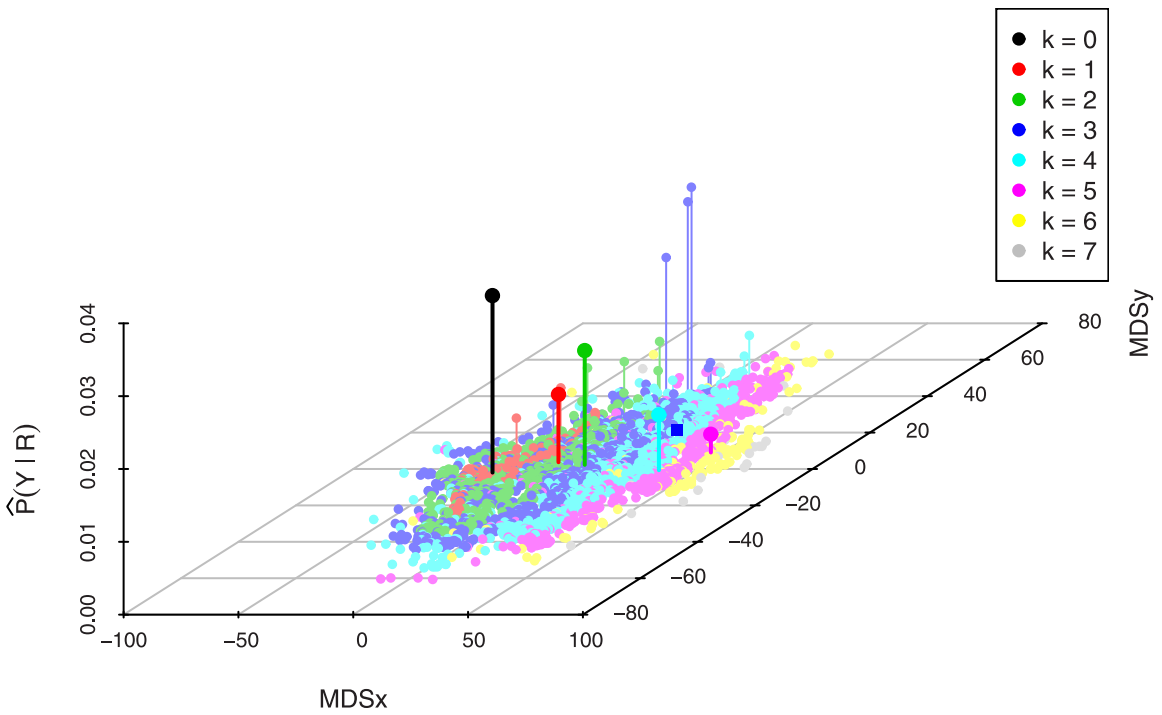


**Figure 8. Estimated posterior distribution of configurations $Y$ based on MCMC samples and projected using multidimensional scaling.** The colors code configurations with different number of binding sites. Bold points mark local centroids, while a square (bold) point highlights the global centroid.
doi:10.1371/journal.pone.0080511.g008

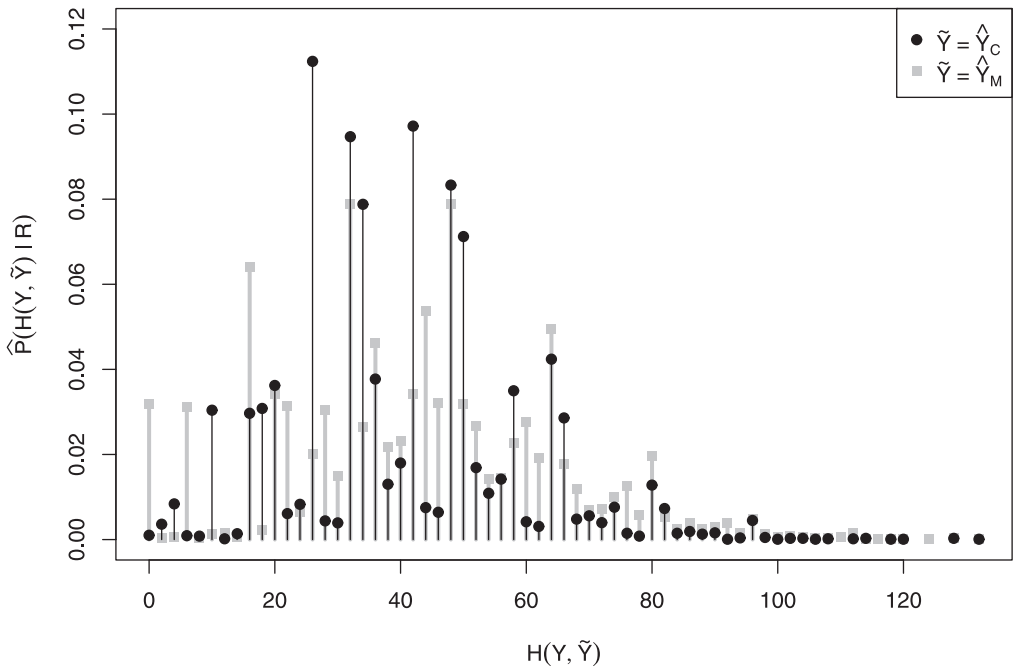


**Figure 9. Estimated posterior distribution of loss function centered at $\tilde{Y}$ for the MAP ($\tilde{Y} = \hat{Y}_M$) and centroid ($\tilde{Y} = \hat{Y}_C$) estimates.**
doi:10.1371/journal.pone.0080511.g009

sequences. More importantly, we discuss and formalize an inferential procedure based on the centroid estimator proposed by Carvalho and Lawrence [20]. As in any Bayesian analysis, we wish to evaluate features of interest in a model based on their posterior distribution; however, if we are required to pick a representative configuration, a point in the parameter space, then a principled approach is to elect a loss function and conduct formal statistical decision analysis. In this sense, by exploring a more refined loss function that depends on position-wise comparisons between sequence states—background or motif positions—we are able to identify a better representative of the posterior space of binding site configurations. Perhaps more importantly, this loss function is meaningful to investigators since is commonly adopted as a metric to measure binding site level accuracy [4,5,25], and so the centroid estimator should be preferred over MAP estimation in principle. Moreover, as pointed out in [20], the centroid estimator better accounts for the distribution of posterior mass; it is more similar to a median than to a mode, and can thus offer better predictive resolution than the MAP estimator [18]. When applied to motif discovery, the centroid estimator captures information in the vicinity of binding site positions through a convolution in marginal posterior distributions of binding sites.

Given the combinatorial number of possible configurations in the parameter space it is not straightforward to identify the centroid estimate through enumeration or even a systematic approach. Yet, we devise an approximative scheme that efficiently optimizes an upper bound on the posterior expected loss and thus provides a related centroid. Despite its heuristic nature, the proposed method has another advantage besides computational convenience: it allows for an informative depiction of the posterior distribution on binding site configurations. First, when defining the local centroids, we are able to assess the contributions from each binding site through their marginal posterior distributions conditional on the number of binding sites, and, in particular, through the convolution of these marginal profiles with the gain filter; secondly, when finding the global centroid we explore the marginal posterior distribution on the number of binding sites. Moreover, other representations might be helpful in understanding the distribution of posterior mass, as in the use of $P_c$ (in Equation 10) to pinpoint the 1-global centroid and measure the overall support of the configurations to a binding site at some specific position in the sequence. These comments are in the spirit of an estimator being also a communicator of the posterior space and the particular choice of prior distribution (see, e.g., Section 4.10 in [24].)

It is important to note that even when the model is accurate, poor inference might fail in recovering relevant features of the space. In Example 2, the MAP estimate is the null configuration, while the centroid indicates three binding sites that represent a group of configurations that jointly pool significant posterior mass. It is also common that the posterior distribution is too complex to be reasonably captured by a single representative; in this case the expected posterior loss could also be used to partition the space and further define additional representatives as conditional estimates on each subspace. This is a direction of work that warrants interest and that we intend to follow next.

Product multinomial and product Dirichlet models are justified as a good working, first approximation based on position independence. There are many extensions to this model that consider DNA strand complementarity [30], a more informative Markov structure for the background composition [31], and an explicit representation of the number of binding sites per sequence [32]. While we adopted a simple hierarchical model to guide the discussion, the proposed methodology is actually broader and the centroid estimators can be obtained from any Bayesian procedure that reports marginal posterior probabilities $\mathbb{P}(c(Y)|R)$ and $\mathbb{P}(Y_k|c(Y),R)$, $k=1,\ldots,c(Y)$, for sequence $R$ and binding site configuration $Y$.

Further improvements can be obtained by specifying a more complex model that accounts, for example, for higher order Markov chains with more states for the background, as in [30,31],

phylogenetic profiles [22], structural information [33], a variable motif length, or dependency among motif positions. As pointed out by Hu, Li, and Kihara [4], motif discovery using sequence only is well known for low signal-to-noise ratio; future extensions would also incorporate other data sources, such as gene expression or ChIP-Seq data, to increase the signal-to-noise ratio. In addition, for future work we intend to extend the model to account for multiple motifs, either from multiple TFs or from a single TF with alternative motifs. While the problem then becomes computationally more challenging, we expect that recursions and estimators similar to the ones discussed here will follow from extra bookkeeping on which motifs are bound at each binding site.

## Supporting Information

**File S1 Derivation of conditional and marginal posterior probabilities for $Y$ and $c(Y)$ and Gibbs sampler for the posterior joint on $Y$ and $\Theta$.** Derivation of conditional posterior probabilities $\mathbb{P}(c(Y)|R,\Theta)$ and marginal posterior probabilities $\mathbb{P}(Y_k|c(Y),R,\Theta)$, along with a routine to compute them in Algorithm 1. A Gibbs sampler to iteratively sample $\Theta|Y,R$ and $Y|\Theta,R$ is given in Algorithm 2.
(PDF)

## Acknowledgments

The author would like to thank Antonio Gomes and Charles Lawrence for the helpful discussions and the reviewers for valuable comments in the text.

## Author Contributions

Conceived and designed the experiments: LC. Performed the experiments: LC. Analyzed the data: LC. Contributed reagents/materials/analysis tools: LC. Wrote the paper: LC.

## References

1. MacIsaac K, Fraenkel E (2006) Practical strategies for discovering regulatory DNA sequence motifs. PLoS Computational Biology 2: e36.
2. GuhaThakurta D (2006) Computational identification of transcriptional regulatory elements in DNA sequence. Nucleic Acids Research 34: 3585–3598.
3. Sandve G, Drablos F (2006) A survey of motif discovery methods in an integrated framework. Biol Direct 1.
4. Hu J, Li B, Kihara D (2005) Limitations and potentials of current motif discovery algorithms. Nucleic Acids Research 33: 4899–4913.
5. Tompa M, Li N, Bailey T, Church G, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology 23: 137–144.
6. Régnier M, Denise A (2004) Rare events and conditional events on random strings. Discrete Mathematics and Theoretical Computer Science 6: 191–214.
7. Pavesi G, Mereghetti P, Mauri G, Pesole G (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Research 32: W199–W203.
8. Stormo G (2000) DNA binding sites: representation and discovery. Bioinformatics 16: 16–23.
9. Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39: 1–38.
10. Lawrence C, Reilly A (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins: Structure, Function, and Bioinformatics 7: 41–51.
11. Bailey T, Elkan C (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning 21: 51–80.
12. Lawrence C, Altschul S, Boguski M, Liu J, Neuwald A, et al. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262: 208–214.
13. Neuwald A, Liu J, Lawrence C (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Science 4: 1618–1632.
14. Liu J, Neuwald A, Lawrence C (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. Journal of the American Statistical Association 90: 1156–1170.
15. Lones MA, Tyrrell AM (2005) The evolutionary computation approach to motif discovery in biological sequences. In: Proceedings of the 2005 workshops on Genetic and evolutionary computation. ACM, pp. 1–11.
16. Lones MA, Tyrrell AM (2007) A co-evolutionary framework for regulatory motif discovery. In: Evolutionary Computation, 2007. CEC 2007. IEEE Congress on. IEEE, pp. 3894–3901.
17. Lones MA, Tyrrell AM (2007) Regulatory motif discovery using a population clustering evolutionary algorithm. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 4: 403–414.
18. Barbieri M, Berger J (2004) Optimal predictive model selection. The Annals of Statistics 32: 870–897.
19. Ding Y, Chan C, Lawrence C (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA 11: 1157–1166.
20. Carvalho L, Lawrence C (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. Proceedings of the National Academy of Sciences of the United States of America 105: 3209–3214.
21. Thompson W, Newberg L, Conlan S, McCue L, Lawrence C (2007) The Gibbs centroid sampler. Nucleic Acids Research 35: W232–W237.
22. Newberg L, Thompson W, Conlan S, Smith T, McCue L, et al. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction. Bioinformatics 23: 1718–1727.
23. Webb-Robertson B, McCue L, Lawrence C (2008) Measuring global credibility with application to local sequence alignment. PLoS Computational Biology 4: e1000077.
24. Berger J (1985) Statistical decision theory and Bayesian analysis. Springer.
25. Pevzner P, Sze S (2000) Combinatorial approaches to finding subtle signals in DNA sequences. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. volume 8, pp. 269–278.
26. Besag J (1986) On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society Series B (Methodological) 48: 259–302.
27. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. Pattern Analysis and Machine Intelligence, IEEE Transactions on 6: 721–741.
28. Liu J (2008) Monte Carlo strategies in scientific computing. Springer Verlag.
29. Gower J (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325–338.
30. Roth F, Hughes J, Estep P, Church G (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nature Biotechnology 16: 939–945.
31. Liu X, Brutlag D, Liu J (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In: Pac Symp Biocomput. volume 6, pp. 127–138.
32. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, et al. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. Journal of Computational Biology 9: 447–464.
33. Xing E, Karp R (2004) MotifPrototyper: a Bayesian profile model for motif families. Proceedings of the National Academy of Sciences of the United States of America 101: 10523.
34. Murrea C, McCawa PS, Baltimorea D (1989) A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and Myc proteins. Cell 56: 777–783.