ABSTRACT

Motivation: The short and degenerate nature of transcription factor (TF) binding sites contributes towards a low signal to noise ratio making it very difficult to separate them from their background. In order to tackle this problem one needs to look at ways of capturing the underlying biophysical properties that best discriminates TF binding sites from their background DNA. One such discriminatory property lies in the observed compositional differences in the nucleotide levels of TF binding sites and background DNA which are a result of processes such as purifying selection and selective preferences of TF binding sites for particular nucleotides or a combination of nucleotides over others.

Results: In this article, we present a hybrid model, referred to as a MonoDi-nucleotide model for robustly detecting TF binding sites. It incorporates both mono- and dinucleotide statistics to optimally partition the base positions of an aligned set of TF binding sites (motif) into a non-redundant sequence of mono and/or dinucleotide segments that maximizes the odds ratio of the binding sites relative to their background DNA. We tested the MonoDi-nucleotide model on the benchmark dataset compiled by Tompa et al. (2005) for assessing computational tools that predict TF binding sites. The performance of the MonoDi-nucleotide model on this data set compares well to, and in many cases exceeds, the performance of existing tools. This is in part attributed to the significant role played by dinucleotides in discriminating TF binding sites from their background DNA. One such discriminatory property lies in the observed compositional differences in the nucleotide levels of TF binding sites and background DNA which are a result of processes such as purifying selection and selective preferences of TF binding sites for particular nucleotides or a combination of nucleotides over others. We tested the MonoDi-nucleotide model on the benchmark dataset compiled by Tompa et al. (2005) for assessing computational tools that predict TF binding sites. The performance of the MonoDi-nucleotide model on this data set compares well to, and in many cases exceeds, the performance of existing tools. This is in part attributed to the significant role played by dinucleotides in discriminating TF binding sites from their background DNA. One such discriminatory property lies in the observed compositional differences in the nucleotide levels of TF binding sites and background DNA which are a result of processes such as purifying selection and selective preferences of TF binding sites for particular nucleotides or a combination of nucleotides over others.

Availability: A Matlab implementation of the MonoDi-nucleotide model can be found at http://www.utoronto.ca/zhanglab/MonoDi/.

Contact: sumedha@cantab.net, Zhaolei.Zhang@utoronto.ca

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Transcriptional regulation is an important means of controlling gene expression in humans and other eukaryotic organisms. Essential to transcriptional regulation is the binding of various modulator transcription factors (TF) onto cis-regulatory elements (binding sites) in the vicinity of the regulated gene. Characterizing these cis-regulatory elements remain an important but difficult task, made difficult by the short (on average ≤ 10 bp) and highly degenerate nature of these sequences.

*To whom correspondence should be addressed.
likely that such clashes, at critical points along the promoter, interfere in the DNAs ability to interact with the binding protein, and hence modulate the efficiency in which the protein binds to the DNA. Evidence of this has been shown by Nussinov (1984), in terms of the local structure of the DNA in the analysis of E.Coli polymerase binding sites. The binding of protein to DNA causes a large number of distortions to the regular twist of the double helix resulting in a departure of the DNA from the classic linear B-DNA to a conformation with a distorted axis which curves towards the protein's recognition helix. The ability of the DNA to tolerate such structural distortions from the classical B-form will, at least partly, determine how well the DNA binds to the protein. Such structural constraints are bound to impose implicit nucleotide dependencies on the DNA. As Calladine and Dickerson (Dickerson, 1983) pointed out, the structural conformation of the DNA is very much dependent on base steps rather than isolated base pairs. Further more, every step in which a steric clash occur, it is projected to its neighboring base pair steps.

Other reasons for associations between nucleotide positions to arise may include processes such as DNA methylation, purifying selection, DNA replication and repair mechanisms and biases in DNA modification (Gentles et al., 2001). Further more, the authors have carried out a study that shows significant differences in the dinucleotide content between TF binding sites and background DNA (see Supplementary Note 1) which provides a compelling reason to look beyond the base independence assumption when discriminating TF binding sites from background DNA. We can express the degree of discrimination of a motif (a collection of fixed length aligned putative binding sites) from its background as an odds ratio between the motif and background models. The odds ratio quantifies the extent to which the motif model derived from its putative binding sites differ from the a priori probabilities of the individual bases of those sites. It can be seen as a statistical measure of the quality of the motif.

In this article we describe a hybrid model referred to as the MonoDi-nucleotide model that discriminates TF binding sites based on their odds ratios. The MonoDi-nucleotide model maximizes the odds ratio of a set of binding sites by selecting an optimal, complete and non-redundant (disjoint) partitioning of the base positions of the sites, into base independent and non-independent segments. This partitioning describes the statistically, independent and non-independent positions of a TF binding site that maximizes its odds ratio over the background. The MonoDi-nucleotide model works in the principle that if dinucleotide dependencies are present in a set of binding sites the model identifies and accommodates it. In the absence of such dependencies, the model only works on mononucleotides. It is a flexible model whose topology adapts to the best discriminatory topology of the binding sites based on mono and/or dinucleotide combinations of nucleotide positions of the sites. As a de novo prediction tool of regulatory elements, the raw data does not provide any prior knowledge of the presence or absence of any dinucleotide dependencies in the sites or their locations within the sites. The MonoDi-nucleotide model also takes advantage of the differences that exist between the non-independent dinucleotide distributions of TF binding sites and background DNA to home in on TF binding sites (see Supplementary Note 4).

We tested the MonoDi-nucleotide model on the benchmark dataset compiled by Tompa et al. (Tompa et al., 2005) for assessing computational tools that predict TF binding sites. The performance of the MonoDi-nucleotide model on this data set compares well or exceeds the performance of existing tools. This is in part attributed to the significant role played by dinucleotides in discriminating TF binding sites from background DNA (see Supplementary Note 4).

2 METHODOLOGY

2.1 Motif model

The motif model is a hybrid model that incorporates both mono- and dinucleotide statistics to optimally partition the base positions of a set of TF binding sites into a non-redundant sequence of mono and/or dinucleotide segments that maximize the odds ratio of the binding sites relative to the background nucleotide distribution. The derivation of the motif model is described below.

2.1.1 Mononucleotide model

In the mononucleotide model, bases of a DNA sequence are assumed to exist independently of each other. This enables the probability of the sequence to be expressed as a product of the probabilities of its individual bases.

**DEFINITION 2.1.** Let $S = \{x_1, x_2, \ldots, x_L\}$, be a nucleotide sequence of length $L$.

In a mononucleotide model, the probability of $S$, $P(S) = p(x_1) p(x_2) \ldots p(x_L)$.

If $S$ is a TF binding site, we can compute its odds ratio, $O(S)$, relative to the background nucleotide distribution by dividing its probability, $P(S|\text{binding site model})$, here forth referred to as $P(S)$, by the product of the background probability of its individual bases, $P(S|\text{background model})$, here forth referred to as $Q(S)$. The background model is simply the a priori probabilities of the individual bases. It is computed from the mono- and di-nucleotide counts of the input sequences after removing the sub-sequences of the motif.

$$O(S) = \frac{P(S)}{Q(S)} = \frac{p(x_1) \cdots p(x_L)}{q(x_1) \cdots q(x_L)}$$

**DEFINITION 2.2.** Let $S = \{ (x_1^{i_1}, x_2^{i_2}, \ldots, x_L^{i_L}) \}$ be a motif of length $L$ of the $N$ aligned sites, $\{S_1, \ldots, S_N\}$, where $x_i$ refers to the $i$th nucleotide of the $j$th sequence of motif $S$.

**DEFINITION 2.3.** Let $S = \{ S_1, \ldots, S_N \}$, be any arbitrary sub-alignment, $S[I] = \{ (x_1^{i_1}, \ldots, x_L^{i_L}) \}$, from position $L$ to position $I$ of the $N$ sequences, $\{S_1, \ldots, S_N\}$, of motif $S$.

If $S$ is a sub-alignment of base independent sub-sequences of $S$, we can compute the normalized odds ratio, $\overline{O}^n(S)$, of $S$ as follows:

$$\overline{O}^n(S) = \left( \prod_{j=1}^{N} O(S_j) \right)^{1/N}$$

$$= \prod_{j=1}^{N} \prod_{i=1}^{L} \left( \frac{p(x_i^{j})}{q(x_i^{j})} \right)^{1/N}$$

$$= \prod_{j=1}^{N} \prod_{i=1}^{L} \left( \frac{p(x_i^{j} = \delta)}{q(x_i^{j} = \delta)} \right)^{1/N}$$

This can be used as a refinement measure of the quality of the motif model.
where \( n'_i \) is the count of the number of nucleotides equal to \( \delta \) along position \( i \) of the \( N \) sequences of \( \text{S} \). The fraction \( \frac{n'_i}{N} \) is the probability of seeing nucleotide \( \delta \) in position \( i \), i.e., \( p(x_i = \delta) \).

We can convert the normalized odds ratio to a normalized log odds ratio, \( I^d \), giving:

\[
I^d(\tilde{S}) = \sum_{i=1}^{T} \sum_{\delta_1, \delta_2 = A}^{T} p(x_i = \delta_1, x_{i+1} = \delta_2) \log_2 \left[ \frac{p(x_i = \delta_1, x_{i+1} = \delta_2)}{q(x_i = \delta_1, x_{i+1} = \delta_2)} \right]
\]

This is the ‘relative entropy’ of \( \tilde{S} \). In information theory, the inner summation is referred to as the ‘Kullback-Leibler distance’.

The probability \( p(x_i = \delta) \) can be computed directly from the available binding sites, as the fraction \( \frac{n'_i}{N} \). In many cases though, the available data does not provide a full picture of the actual nucleotide distributions of the sites. The general means of circumventing this problem is to add pseudo counts to account for unseen instances of the data. Hertz and Stormo (Hertz et al., 1999) proposed adding pseudo counts proportional to the background distribution of the nucleotide bases, which is the method we follow in this article, therefore:

\[
p(x_i = \delta) = \left[ n'_i + q(x_i = \delta) \right] / [N + 1]
\]

where \( \sum_{\delta = A}^{T} q(x_i = \delta) = 1 \).

### 2.1.2 Dinucleotide model

In the dinucleotide model, we assume that nucleotide correlations do occur, however due to computational complexities involved with modelling higher order nucleotide correlations in TF binding sites we only consider correlations between neighboring base pairs. Under this model, the probability of a sequence \( S \) can be expressed as:

\[
P(S) = p(x_1, \ldots, x_L) = p(x_1, x_2) p(x_2, x_3) \cdots p(x_{L-1}, x_L)/p(x_2) \cdots p(x_{L-1})
\]

If \( \tilde{S}(t < f) \) is a sub-alignment of \( S \) whose nucleotide positions are correlated between neighboring base pairs, we can compute its normalized odds ratio, \( I^d \), using the dinucleotide model as follows:

\[
I^d(\tilde{S}) = \sum_{i=1}^{T} \sum_{\delta_1, \delta_2 = A}^{T} p(x_i = \delta_1, x_{i+1} = \delta_2) \log_2 \left[ \frac{p(x_i = \delta_1, x_{i+1} = \delta_2)}{q(x_i = \delta_1, x_{i+1} = \delta_2)} \right]
\]

where \( n'_i, \delta_1, \delta_2 \) is the count of the number of dinucleotides that are equal to \( \delta_1 \) in position \( i \) and \( \delta_2 \) in position \( i+1 \) of the \( N \) sequences of \( \text{S} \). The fraction \( \frac{n'_i, \delta_1, \delta_2}{N} \) is the probability of seeing nucleotide \( \delta_1 \) in position \( i \) and \( \delta_2 \) in position \( i+1 \), i.e., \( p(x_i = \delta_1, x_{i+1} = \delta_2) \).

Like the mononucleotide model, we can compute the normalized log odds ratio, \( I^d(\tilde{S}) \), for the dinucleotide model giving:

\[
I^d(\tilde{S}) = \sum_{i=1}^{T} \sum_{\delta_1, \delta_2 = A}^{T} p(x_i = \delta_1, x_{i+1} = \delta_2) \log_2 \left[ \frac{p(x_i = \delta_1, x_{i+1} = \delta_2)}{q(x_i = \delta_1, x_{i+1} = \delta_2)} \right]
\]

If an independent background model is assumed, the joint background probability \( q(x_i, x_{i+1}) \) can be replaced by product \( q(x_i) q(x_{i+1}) \). Similar to the mononucleotide model, pseudo counts are added when computing the probability \( p(x_i = \delta_1, x_{i+1} = \delta_2) \) giving:

\[
p(x_i = \delta_1, x_{i+1} = \delta_2) = \frac{n'_i, \delta_1, \delta_2 + q(x_i = \delta_1, x_{i+1} = \delta_2)}{N + 1}
\]

where \( \sum_{\delta_1, \delta_2 = A}^{T} q(x_i = \delta_1, x_{i+1} = \delta_2) = 1 \).

In order to account for the difference in the degrees of freedom between the more complex dinucleotide model and the equivalent less complex mononucleotide model we scaled down the dinucleotide model (at the dinucleotide level) by subtracting a factor \( \lambda \) from the normalized log odds ratio (see Supplementary Note 2). We set \( \lambda = \log_2(1.23) \) (derived from Gentles et al., 2001) which ensures at least a 1.23-fold increase in odds of a dinucleotide under the dinucleotide model relative to an equivalent mononucleotide model.

### 2.1.3 MonoDi-Nucleotide model

Transcription factor binding sites can have a combination of both base independent and non-independent segments of nucleotide positions. The odds ratio of such sequences can be computed as a product of the odd ratios of the individual segments computed using the appropriate model.

**Definition 2.4.** Let \( M = \{ \tilde{S}_{1}^{m_1}, \ldots, \tilde{S}_{1}^{m_L} \} \) be a set of, \( M \), non-overlapping base independent sub-alignments of \( S \). \( \tilde{S}_{1}^{m_1} = \{ x_{i_1}^1, \ldots, x_{i_1}^{m_1} \}, \ldots, \{ x_{i_L}^1, \ldots, x_{i_L}^{m_L} \} \) where \( 1 \leq i_1 \leq \cdots \leq i_L, l_1 \leq \cdots \leq l_L \leq L \).

**Definition 2.5.** Let \( D = \{ \tilde{S}_{2}^{d_1}, \ldots, \tilde{S}_{2}^{d_L} \} \) be a set of, \( D \), non-overlapping base dependent (between neighboring bases) sub-alignments of \( S \). \( \tilde{S}_{2}^{d_1} = \{ x_{i_1}^1, \ldots, x_{i_1}^{d_1} \}, \ldots, \{ x_{i_L}^1, \ldots, x_{i_L}^{d_L} \} \) where \( 1 \leq i_1 \leq \cdots \leq i_L \leq L \).

We can compute the odds ratio of the union of the two sets. It is simply the product of the odd ratios of the individual sets i.e.

\[
O(M \cup D) = \prod_{i=1}^{M} O^i(\tilde{S}_{1}^{m_i}) \prod_{j=1}^{D} O^j(\tilde{S}_{2}^{d_i})
\]

The normalized log odds ratio of the combined sets, is given by

\[
I(M \cup D) = \sum_{i=1}^{M} I^i(\tilde{S}_{1}^{m_i}) + \sum_{j=1}^{D} I^j(\tilde{S}_{2}^{d_j})
\]

The MonoDi-nucleotide model enables us to combine the odd ratios of base independent and non-independent nucleotide sequences.
Each position of the binding sites (column of the alignment) maps to a single node in Layer 1 of the MonoDi-nucleotide network such that the first node maps to the first column, the second node, to the second column, and so on.

\[ n_i^1 \rightarrow S[i] \quad \text{for} \quad i = 1, \ldots, L \]

Hence, the number of nodes in Layer 1 of the network is equal to the length of the binding sites (\( L \) nodes for \( L \) nucleotide positions). These nodes represent the normalized log odds ratios of the nucleotides of the columns they map to. Nodes of Layer 1 are ordered monotonically in the order of the nucleotide positions they map to. Within Layer 1, nodes are connected by directed edges, going from left to right, between neighboring nodes

\[ n_i^1 \rightarrow n_i^{1+1} \quad \text{for} \quad i = 1, \ldots, L - 1 \]

Nodes of Layer 2 map to dinucleotides with a node for each positional pair (adjacent columns of the alignment), giving \( L - 1 \) nodes for \( L \) nucleotide positions.

\[ n_i^2 \rightarrow S[i+1] \quad \text{for} \quad i = 1, \ldots, L - 1 \]

Positions are taken in successive overlapping pairs (i.e. pairs of the form \([1, 2], [2, 3], \ldots, [L-1, L]\)). Nodes of Layer 2 represent the normalized log odds ratios of the dinucleotides of the adjacent pairs of columns they map to. Like Layer 1, Layer 2 is also ordered monotonically in the order of the first nucleotide position in the pair of nucleotide positions they map to. Nodes of Layer 2 are connected to each other only through a unique node in Layer 3. This node in Layer 3 represents the negative normalized log odds ratios of the intersecting column of the two, Layer 2 nodes that it connects.

\[ n_i^2 \rightarrow n_i^3 \quad \text{for} \quad i = 1, \ldots, L - 1 \]

\[ n_i^2 \rightarrow n_i^{2+1} \quad \text{for} \quad i = 1, \ldots, L - 2 \]

Nodes of Layer 1 connect to nodes of Layer 2 whose first position is the immediate successor to its position. Likewise, nodes of Layer 2 connect to nodes of Layer 1 whose position is the immediate successor to its second position.

\[ n_i^1 \rightarrow n_i^2 \quad \text{for} \quad i = 1, \ldots, L - 1 \]

\[ n_i^2 \rightarrow n_i^{2+2} \quad \text{for} \quad i = 1, \ldots, L - 2 \]

These connections enable the network to alternate between the mononucleotide and dinucleotide models.

Finally, the Start node of the network connects to the first node of Layer 1 and the first node of Layer 2. The last node of Layer 1 and Layer 2 connects to the End node of the network.

\[ n_i^0_{\text{Start}} \rightarrow n_i^1; \quad n_i^0_{\text{End}} \rightarrow n_i^2; \quad n_i^0_{\text{End}} \rightarrow n_i^0_{\text{End}} \]

We can use the MonoDi-nucleotide network to obtain an optimal partitioning of the nucleotide positions. In order to do this, we use a dynamic programming algorithm to systematically fill the MonoDi-nucleotide network. It proceeds as follows:

**Definition 2.9.** Let \( C(n_i^k), i \in [1, \ldots, \text{End}] \) and \( k \in \{0, 1, 2, 3\} \), be the maximum cumulative normalized log odds ratio, along the optimum path, at node \( n_i^k \), starting from the Start node \( n_i^0_{\text{Start}} \):

- We initialize the network by setting \( C(n_i^0_{\text{Start}}) = 0 \).
For any other node \( n_i^k \), for valid \( i \in \{1, \ldots, \text{End} \} \) and \( k \in \{0, 1, 2, 3\} \), of the network, we have
\[
C(n_i^k) = \tilde{T}(n_i^k) + \max\{C(\text{Parent}(n_i^k))\}
\]
where
\[
\tilde{T}(n_i^k) = \begin{cases} 
I^m(S[i]) & \text{if } k = 1; \\
I^l(S[i+1]) & \text{if } k = 2; \\
-I^m(S[i+1]) & \text{if } k = 3.
\end{cases}
\]

The function \( \text{Parent}(n_i^k) \) returns the parent nodes of \( n_i^k \) (i.e., the nodes that connect to \( n_i^k \)).

- We keep track of the direction of the optimal path using the function \( D(n_i^k) \) defined as
\[
D(n_i^k) = \max\{C(\text{Parent}(n_i^k))\} \quad \text{for} \quad i = 1, \ldots, \text{End}
\]
\[
k \in \{0, 1, 2, 3\}
\]
(In case there is more than one path, we select the path along the mononucleotide model giving it priority over the dinucleotide model.)

- The optimal partition, \( \hat{P}(S) \), of the nucleotide positions of \( S \) is obtained by tracing back the path from the \( \text{End} \) node to the \( \text{Start} \) node.
- The optimal normalized log odds ratio, \( \hat{S}(S) \), of motif \( S \) partitioned by \( \hat{P}(S) \) is given by \( C(n_{\text{End}}^0) \).

The network can be completed by traversing the nodes in the order
\[
\mathbf{R} = n_{\text{Start}}^0, n_1^1, n_2^1, n_3^1, n_1^2, n_2^2, n_3^2, \ldots, n_{\text{End}}^0
\]
For the benefit of the readers, an illustrative example is provided in Supplementary Material Worked Example.

### 2.3 Motif significance score

The motif significance (MS) score is used to rank motifs. The MS score, \( E \), of a motif \( S \) with an optimal normalized log odds ratio of \( \hat{S}(S) \) is given by
\[
E(S) = \hat{S}(S) + w \times \log_2[(N/T)/(1 - N/T)]
\]
where \( N \) is the number of sites in motif \( S \) and \( T \), the total number of possible sites given the sequences. The MS score incorporates the posterior probability of sampling motif \( S \). This is similar to the posterior probability of a model defined in Neuwald et al. (1995). However, we do not add weighted pseudosites to the actual number of sites but instead weight this probability by a fraction \( w \in (0,1) \). The default setting for \( w \) is 0.8. The MS score described here does not include a term for variable motif lengths as only fixed length motifs are considered.

### 2.4 Sampling the sequence space

Computational techniques such as Gibbs sampling that rely on randomly sampling a given sequence space for finding statistically significant motifs depend heavily on the accuracy and efficiency of the particular sampling strategy used. It is generally done uniformly over the given sequence space. This is however an inefficient approach that exacerbates as the number and, more so, the length of the sequences being sampled increase.

We have observed that the placement of TF binding sites in the upstream regions of co-regulated genes for some transcription factors tend to locate within relatively similar distances from the transcription start site of those genes regardless of the absolute distance from it (see Supplementary Note 3). It is an observation relative to the length of the upstream region considered. This is illustrated in Figure 2. Given this observation, sampling the sequence space can be preformed more efficiently if it is concentrated more on the regions where TF binding sites tend to locate over regions where they are less likely to be found instead of uniformly scanning the entire space. However, sampling should not be prejudicial to any particular spread of TF binding sites.

If we assume that the probability of a TF binding site, \( s \), being at a distance, \( l \), from the transcription start site is proportional to its placement on the sequence which is given by the distance to \( s \) from either end of the sequence i.e \( l \) and \( L-l \), we can express the probability of \( s \) being located at a distance \( l \) from the transcription start site as a function of the distances \( l \) and \( L-l \) as \( P(l,s) = k(l^p/(L-l)^q) \), where \( L \) is the length of the sequence and \( k \), a normalizing constant given by \( k = \int p^l(L-l)^q \).

This distribution can be converted to a standard beta distribution by normalizing the distances \( l \) and \( L-l \) to lie in the range \((0,1)\) by dividing by \(L+1\) and replacing the parameters \( p \) and \( q \) by \( p-1 \) and \( q-1 \), respectively (where \( p > 0, q > 0 \)). The parameters \( p \) and \( q \), or more precisely, the location parameter \( \alpha = p/(p+q) \) and the scale parameter \( \beta = p+q \), determines the shape of the distribution. Given that parameters \( \alpha \) and \( \beta \) describe the distribution of the placement of binding sites from the transcription start site we can use this distribution to sample the sequence space much more efficiently than sampling it uniformly. For a given set of observed distances, we can compute the ML estimates of parameters \( \alpha \) and \( \beta \) very efficiently (see Hahn et al., 1994; Scarborough, 1955).

The sequence space is sampled using a dynamic beta distribution, \( \hat{B}(\hat{\alpha}, \hat{\beta}) \), which at any given time represents the distribution of distances of the sites in the motif with maximum MS score found to that point. We start by computing the ML estimates of \( \alpha \) and \( \beta \) representing the distribution that fits the
observed distances of the sites in the initial motif (one can also start with the uniform distribution by setting $\alpha = 0.5$ and $\beta = 2$). This distribution is used to sample the sequence space until a new motif is found with greater MS score. The distances of the sites in this motif is used to re-compute $\alpha$ and $\beta$, and the resulting distribution is used to sample the sequence space. This process continues until the final motif is obtained. If we start with a uniform sampling distribution, the dynamic update step of the sampling distribution can be delayed until a good initial motif configuration is found (e.g. a motif configuration with a positive MS score).

2.5 Motif Sampler
Motifs are sampled using a simulated annealing based Gibbs sampling algorithm. The algorithm is fully described in Supplementary Algorithm 1. The algorithm updates the motif configuration by visiting sites randomly according to the proposal distribution described in Section 2.4. Motifs are scored using the motif model described in Section 2.1.

3 RESULTS AND DISCUSSION
In 2005, Tompa et al. introduced a collection of datasets for assessing the performance of computational tools that predict transcription factor binding sites. These datasets comprised of eukaryotic binding sites belonging to 52 transcription factors representing four different species, 6 belonging to fly, 26 belonging to human, 12 belonging to mouse and 8 belonging to yeast. The binding sites of each transcription factor were presented in three different background models, ‘real’, ‘generic’ and ‘Markov’. For each of the different type of background model, the known positions and orientations of the binding sites were kept unchanged. Four additional datasets of type ‘Markov’ containing no planted binding sites were included as negative controls.

Using these datasets Tompa et al. evaluated the performance of 13 different computational tools for de novo prediction of regulatory elements. We evaluated the MonoDi nucleotide model on these datasets according to the performance measures described in Tompa et al. and compared it to the 13 tools evaluated in that paper. The comparative results of this evaluation are described below. More details of the datasets and the different statistics used to assess tool performance quality can be found in Tompa et al. (2005).

Figure 3 shows the comparative results for the seven statistics (see Supplementary Materials for details), nucleotide-level sensitivity ($n$Sn), nucleotide-level positive predictive value ($n$PPV), nucleotide-level performance coefficient ($n$PC), nucleotide-level correlation coefficient ($n$CC), site-level sensitivity ($s$Sn), site-level positive predictive value ($s$PPV) and site-level average site performance ($s$ASP)—summarized over all 56 datasets (regardless of species or background model). The MonoDi nucleotide model outperforms the other tools in all indicators except for the $x$PPV in which it falls behind Weeder. However, as explained in Tompa et al. (2005), the $x$PPV values tend to be exaggerated for those programs that make no predictions on datasets. The MonoDi-nucleotide model had a motif prediction on every dataset it attempted a prediction while Weeder on the other hand had 17 datasets which it predicted no motif (see Supplementary Table 1) giving it a biased advantage on the $x$PPV scale.

Figure 4 gives the breakdown of the performance, highlighted by the correlation coefficient, $n$CC, of each tool on the datasets of the different species (regardless of the background model). From Figure 4, we can see that the MonoDi-nucleotide model does better than the other tools in all species except for yeast. The three tools, Weeder, MotifSampler and MEME do better than the MonoDi-nucleotide model on the yeast datasets. We believe that the reason that Mono-di model works better for fly, mouse and human and less so for yeast is because the yeast genome is less biased in di nucleotide frequencies (Gentles et al., 2001).

Figure 5 breaks down the datasets according to the different background models, ‘real’, ‘generic’ and ‘Markov’ (regardless of the species). The MonoDi-nucleotide model performs better than the other tools in predicting motifs in the ‘real’ and ‘Markov’ background models. It does better than the other tools for the ‘generic’ background model as well, except for Weeder which performance slightly better.

Further comparative evaluations of the MonoDi-nucleotide model is presented in the Supplementary Material. Supplementary Figure 1 shows the comparative performance of the MonoDi-nucleotide model on the datasets with real background models as measured by the seven statistics of Figure 3. Supplementary Figure 2 replicates the same seven statistics for the analyzed tools over just the datasets with generic and Markov background models. Supplementary Figure 3 provides an idea of the performance of the individual tools when considered in absence of the yeast datasets. In all these evaluations, we see that the MonoDi-nucleotide model outperforms the other tools.

Significant motifs of the MonoDi-nucleotide model were selected based on their MS score. We considered motifs with an MS score cutoff of 4.5 or greater for the above evaluation. The value 4.5 was selected because it yielded comparatively similar amounts of predicted motifs as the other tools analyzed.
Figure 3 and others. The MonoDi-nucleotide model uses a simple construction that extends the well established mono-nucleotide weight matrix model to incorporate dinucleotide information from both the motif and background. The MonoDi-nucleotide model described here does ‘de novo’ prediction of binding sites, but can be easily adapted with training data to scan for binding sites of known transcription factors.

Identifying TF binding sites is a challenging task, due mainly to our lack of understanding of the true nature of the signal(s) that discriminate these sites from background DNA. It is evident that models such as position specific weight matrices are too restrictive in many instances in representing the degenerate nature of TF binding sites. In order to better discriminate TF binding sites, we need to look at more subtle signals that distinguish these sites from background DNA. The relative difference in dinucleotide content that manifest as strong nucleotide associations in TF binding sites and background DNA is one such signal which the MonoDi-nucleotide model exploits when looking for TF binding sites. Discriminative signals such as nucleotide associations are not universal and hence are not evident in the binding sites of all transcription factors. Neither are they homogeneous in the nature of their occurrence across TF binding sites of different transcription factors. This is why it is necessary to develop flexible models such as the MonoDi-nucleotide model whose topology can adapt to the best discriminatory topology of the binding sites based on the optimum partitioning of the base positions of the sites into independent mono nucleotide and/or correlated dinucleotide segments.

4 CONCLUSION

We have introduced the a new model for detecting TF binding sites based on an optimal combination of mono- and/or di-nucleotide log likelihood scores. The model outperforms many of the contemporary motif detection tools as revealed in Supplementary Figures 4–9 with the performance indicators of Figures 3–5 and Supplementary Figures 1–3 with the predictions from the MonoDi-nucleotide model taken with a conservative MS cutoff of 5 and a more liberal MS cutoff of 3.5.

Increasing this value increases the quality of the predicted motifs and hence the performance of the MonoDi-nucleotide model however it reduces the number of motifs predicted. Reducing the MS cutoff score on the other hand results in more datasets with predicted motifs but with a less average quality. Supplementary Figures 4–9 shows the performance indicators of Figures 3–5 and Supplementary Figures 1–3 with the correlation coefficient (nCC) by background model.

ACKNOWLEDGEMENTS

This work was supported in part by new faculty start-up funds from the University of Toronto to Z.Z. and grants from Genome Canada through the Ontario Genomics Institute.

Conflict of Interest: none declared.

REFERENCES


