**Discovery and Validation of Information theory-based Transcription Factor and Cofactor Binding Site Motifs**

**Supplementary Methods**

**Ruipeng Lu1, Eliseos J. Mucaki2 and Peter K. Rogan1,2,3,4**

**1 Department of Computer Science, 2 Department of Biochemistry, 3 Department of Oncology, Western University, London, Ontario, N6A 4L6, Canada, 4 Cytognomix Inc., London, Ontario, N5X 3X5, Canada**

1. **Bioinformatic tools**

**1.1) The Bipad algorithm**

**1.2) The Maskminent motif discovery pipeline**

**1.2.1) The Maskminent program versus the Bipad program**

**1.2.2) The flowchart of the pipeline and an example**

**1.2.3) The commands to run the Maskminent program and the Scan program**

**1.2.4) The mathematical formalization of the masking technique**

**1.2.5) The mathematical formalization of selecting top peaks**

**1.2.6) The distributions of** $R\_{i}$ **values of binding sites in top peaks versus in bottom peaks**

**2) Statistical analyses on iPWMs**

**2.1) The derivation on the relationship between binding energy** $\left(log\_{2}K\_{di}\right)$ **and** $R\_{i}$ **values**

**2.2) Evaluation of the linearity between binding energy and** $R\_{i}$ **values using F tests**

**2.3) An example slope analysis**

**1) Bioinformatic tools**

**1.1) The Bipad algorithm**



 Figure 1. A high-level illustration of the Bipad algorithm

**Mathematical principles:**

Given the width $J$ of the iPWM to be derived, a peak dataset can be viewed as a multiple alignment search space $θ $in which each multiple alignment whose size equals the number of peaks in the dataset is generated by taking a sequence fragment of length $J$ from each peak. Given a multiple alignment, the entropy of the position $l \left(1\leq l\leq J\right)$ can be computed from:

$$E\_{l}=\sum\_{b\in B}^{}f\left(b,l\right)log\_{2}\frac{1}{f\left(b,l\right)} , B=\left\{A,C,G,T\right\} [1]$$

where $f\left(b,l\right)$ is the frequency of base $b$ at position $l$.

Bipad (1)uses a Monte Carlo framework to search subspaces in$ θ$, in an attempt to find the globally optimal multiple alignment with the minimum entropy (i.e. the maximum information content) in$ θ$. This process is independent of the nucleotide composition of input sequences. Therefore, the objective function is:

$$oMA=arg\min\_{MA\in θ}\left(E\_{MA}\right) [2], E\_{MA}=\sum\_{l=1}^{J}E\_{l} or\sum\_{s\in \left\{L,R\right\}}^{}\left(\sum\_{l=1}^{J\_{s}}E\_{l}\right) [3]$$

where $oMA$ is the optimal multiple alignment with the minimum entropy, $ MA$ is one contiguous or bipartite multiple alignment in $θ$, $E\_{MA}$ is the entropy of $MA$, $J\_{s}$ is the length of the left or right half site in $MA$.

**1.2) The Maskminent motif discovery pipeline**

**1.2.1) The Maskminent program versus the Bipad program**

The Maskminent programis obtained after adding the masking technique to the Bipad program. Therefore, it is a generalization of Bipad, and is able to reveal additional binding motifs that Bipad is not able to discover from the same ChIP-seq dataset.

**1.2.2) The flowchart of the pipeline and an example**



Figure 2. The flowchart of the Maskminent motif discovery pipeline \*

\* $iPWM\_{N}$ is the initial iPWM derived from all peaks of the dataset. $iPWM\_{200}$ is the iPWM derived from the top 200 peaks. The initial range containing the minimum threshold peak strength that can generate the primary/cofactor motif is from the strength of the 200th peak (i.e. the initial value of $G$) to the strength of the last peak (i.e. the initial value of $S$). $S$ are the smaller bound of the current range containing the minimum threshold, and $G$ is the greater bound (i.e. the current threshold). $iPWM\_{G}$, $iPWM\_{S}$, $iPWM\_{M}$ are respectively the iPWMs derived from the top peaks above $G$, $S$, $M$. $d\left(iPWM\_{G},iPWM\_{M}\right)$ is the Euclidean distance between $iPWM\_{G}$ and $iPWM\_{M}$, and $d\left(iPWM\_{S},iPWM\_{M}\right)$ is the Euclidean distance between $iPWM\_{S}$ and $iPWM\_{M}$. The approximately minimum threshold that we finally obtain is $G$ of the final range.

**Example**: Suppose we want to apply the Maskminent pipeline to a ChIP-seq dataset with 2,000 peaks. The following list shows each step of going through the above flowchart.

1. Derive $iPWM\_{N}$ from all the 2,000 peaks.
2. $iPWM\_{N}$ exhibits a noise motif.
3. Derive an iPWM $iPWM\_{1}$ from all the 2,000 peaks with masking $iPWM\_{N}$.
4. $iPWM\_{1}$ exhibits a cofactor motif. So $iPWM\_{1}$ is output.
5. Derive an iPWM $iPWM\_{2}$ from all the 2,000 peaks with masking $iPWM\_{N}$ and $iPWM\_{1}$.
6. $iPWM\_{2}$ exhibits a noise motif. Next we will use the thresholding technique.
7. Sort all the 2,000 peaks in the descending order of signal strengths.
8. Select the top 200 peaks.
9. Derive an iPWM $iPWM\_{200}$ from the top 200 peaks.
10. $iPWM\_{200}$ exhibits the primary motif. So $iPWM\_{200}$ is output.
11. Assign the strength of the 200th peak to$ G$, and the strength of the 2,000th peak to$ S$. Assign $iPWM\_{200}$ to $iPWM\_{G}$, and $iPWM\_{N}$ to $iPWM\_{S}$.
12. Because 2,000-200>500, assign the strength of the 1,000th peak to $M$. (1,000 = rounding (200+2,000)/2 to the nearest multiple of 500)
13. Select the top 1,000 peaks.
14. Derive $iPWM\_{M}$ from the top 1,000 peaks.
15. $d\left(iPWM\_{G},iPWM\_{M}\right)$ is smaller than $d\left(iPWM\_{S},iPWM\_{M}\right)$. So $iPWM\_{M}$ exhibits the primary motif, and it is output.
16. Assign $M$ (i.e. the strength of the 1,000th peak) to $G$, and $iPWM\_{M}$ to $iPWM\_{G}$.
17. Because 2,000-1,000>500, assign the strength of the 1,500th peak to $M$. (1,500 = rounding (1,000+2,000)/2 to the nearest multiple of 500)
18. Select the top 1,500 peaks.
19. Derive $iPWM\_{M}$ from the top 1,500 peaks.
20. $d\left(iPWM\_{G},iPWM\_{M}\right)$ is smaller than $d\left(iPWM\_{S},iPWM\_{M}\right)$. So $iPWM\_{M}$ exhibits the primary motif, and it is output.
21. Assign $M$ (i.e. the strength of the 1,500th peak) to $G$, and $iPWM\_{M}$ to $iPWM\_{G}$..
22. Because 2,000-1,500=500, the pipeline is stopped.

Finally, the true TF binding motifs we obtain include a cofactor motif (i.e. $iPWM\_{1}$) and the primary motif. The approximately minimum threshold that we finally obtain is the strength of the 1,500th peak, and the most accurate iPWM exhibiting the primary motif that we obtain is derived on the top 1,500 peaks.

**1.2.3) The commands to run the Maskminent program and the Scan program**

In the Maskminent program, the –m option indicates whether to switch on masking, otherwise the commands specifying the run to build contiguous or bipartite iPWMs are similar to Bipad, respectively:

./Maskminent –n LengthOfiPWM –y NumberOfCycles –f ChIPseqFile [-m MaskFile] [-2]

./Maskminent –l LengthOfLeftHalfSite –r LengthOfRightHalfSite –a MinSpacerSize –b MaxSpacerSize -y NumberOfCycles –f ChIPseqFile –d/i

The command to run the Scan program is:

./Scan iPWMFile SequenceFile (1 LengthOfiPWM)/(2 LengthOfHalfSite MinSpacerSize MaxSpacerSize)

The ChIPseqFile consists of a set of DNA sequences, each of which may differ in length, as separate consecutive records.

**1.2.4) The mathematical formalization of the masking technique**

The masking technique is implemented by scanning a dataset with prior iPWMs, then recording the coordinates of all the predicted binding sites, and skipping these intervals when reanalyzing the dataset. This reduces the entire multiple alignment search space by the predicted sites of the preceding motifs. So the objective function of the Bipad algorithm (i.e. the Equation 2) changes accordingly:

$$oMA=arg\min\_{MA\in θ-θ\_{i}}\left(E\_{MA}\right) [4]$$

where $θ\_{i}$ is the search space covered by the predicted binding sites of the previous iPWMs to be masked.

**1.2.5) The mathematical formalization of selecting top peaks**

Suppose that we want to select the top n peaks, after all peaks in a ChIP-seq dataset *D* are ranked in the descending order of signal strengths. *size(D)* is the number of peaks in *D*, and *min(D)* and *max(D)* are the lowest and highest signal values among all peaks in D, respectively. Then, the operation of thresholding the dataset *D* to obtain the top *n* peaks is defined by:

threshold(*D, n*) = $D\_{n}, where \left(D\_{n}⊂D\right)∧\left(size\left(D\_{n}\right)=n\right)∧\left(min\left(D\_{n}\right)\geq max\left(D-D\_{n}\right)\right)$, *n < size(D)* [5]

The masking and thresholding techniques decrease the multiple alignment search space from distinct directions, as masking removing a portion in which the prior iPWMs are contained, and thresholding increasing the possibility of a non-noise motif possessing the smallest entropy in the remaining search space.

**1.2.6) The distributions of** $R\_{i}$ **values of binding sites in top peaks versus in bottom peaks**

At the end of the thresholding process we obtain the approximately minimum threshold peak strength. The peak set above this threshold contains the maximum number of top peaks that can produce the primary/cofactor motif. However, the weak peaks below this threshold do not necessarily contain weak or missing binding sites. In fact, the distribution of *Ri* values of binding sites in these bottom peaks is similar to that in the top peaks used to derive the primary motif. The fact that inclusion of the bottom peaks results in the disappearance of the primary motif implies that overall the binding sites in the top peaks have higher conservation levels (i.e. information contents). Below an example on the RUNX3 ChIP-seq dataset is given.

We took the greatest *Ri* value in each peak after using the iPWM derived from the top 64,500 peaks to scan each peak in this dataset. The following graph shows the distributions of *Ri* values for the top 64,500 peaks versus for the 3,465 bottom peaks.



**2) Statistical analyses on iPWMs**

**2.1) The derivation on the relationship between binding energy** $\left(log\_{2}K\_{di}\right)$ **and** $R\_{i}$ **values**

We found that the *Ri* values are related to binding energy (or affinity). The frequencies of binding sites appearing in a ChIP-seq dataset are related to their binding energy (*log2Kd*), which is delineated by Equation 6 derived by us starting from Levine et al. (2). And the *Ri* values of binding sites are proportionate to their frequencies, since the stronger a binding site is, the more the number of times it is bound by the TF will be in a ChIP-seq assay.

$K\_{di}≈ \frac{F\_{1}^{'}K\_{d1}}{F\_{i}^{'}} [6]$

where $K\_{di}$ is the dissociation constant of the $i$th predicted binding site, $F\_{i}^{'}$ is the frequency of the $i$th site in a round of bounding (i.e. the frequency of the $i$th site appearing in the ChIP-seq dataset), $K\_{d1}$ and $F\_{1}^{'}$ are these values for the strongest site (i.e. the consensus sequence).

The Equation 6 was derived by recognizing that the thermodynamics of a population of TF-bound sequences is similar to a SELEX experimental framework (3). Below we give the detailed derivation.

From Levine et al. (2), first we obtain Equation 7:

$$\frac{F\_{i}^{'}}{F\_{1}^{'}}=\frac{K\_{d1}+\left[Tf\right]}{K\_{di}+\left[Tf\right]}⋅\frac{F\_{i}}{F\_{1}} [7]$$

where $F\_{i}$ is the frequency of the $i$th site in the prior round of bounding (i.e. in the genome), $F\_{1}$ is this value for the strongest site (i.e. the consensus sequence), $\left[Tf\right]$ is the concentration of unbound TF.

Multiply both sides of Equation 7 by $F\_{1}^{'}$, then we obtain Equation 8:

$$F\_{i}^{'}=\frac{K\_{d1}+\left[Tf\right]}{K\_{di}+\left[Tf\right]}⋅\frac{F\_{i}}{F\_{1}}⋅F\_{1}^{'} [8]$$

Given that the consensus sequence is an extremely infrequent binding site both in the unselected population of binding sites and in the genome (3), we assume that its frequency will be similar during the early rounds of selection (i.e. $F\_{1}=F\_{1}^{'}$). Then, we obtain Equation 9:

$$F\_{i}^{'}=\frac{K\_{d1}+\left[Tf\right]}{K\_{di}+\left[Tf\right]}⋅F\_{1}^{'} [9]$$

Solve for $K\_{di}$, then we obtain Equation 10.

$$K\_{di}= \frac{F\_{1}^{'}\left(K\_{d1}+[Tf]\right)}{F\_{i}^{'}}-\left[Tf\right] [10]$$

$\left[Tf\right]$ is negligible for most TFs, because the concentrations of TFs inside cells are quite low (i.e. in the nM range) and $\left[Tf\right]$ is only a fraction of them (4). Therefore, we obtain Equation 6 which suggests that the dissociation constants of binding sites are inversely proportional to their frequencies.

**2.2) Evaluation of the relationship between binding energy and** $R\_{i}$ **values using F tests**

Higher $R\_{i}$ values have lower binding energy. Primary/cofactor motifs are expected to demonstrate this relationship, whereas noise motifs are not. F tests can be used to evaluate this relationship, because they measure whether the linear regression model between the two variables provides a significantly better fit than the ‘nested’ constant regression model. A large F value means that the linear regression model has an slope well below 0, so primary/cofactor motifs are expected to have greater F values than noise motifs.

The frequency of a binding site appearing in the ChIP-seq dataset was calculated as follows: if only one binding site appeared in a peak, the frequency of this site would increase by one. In the case that a peak contained multiple binding sites, the frequency of each site would gain a fraction of one based on the proportion of its *Ri*value accounting for the sum of *Ri*values of all the sites.

For all the TFs we used 10-7M as the default value for the dissociation constant $K\_{d1}$. This is because on an iPWM using different values for $K\_{d1}$ will lead to the same F value. The detailed proof is given below.

Assume that for $K\_{d1}$ we use two different values $K\_{d1}^{a}$ and $K\_{d1}^{b}$. According to Equation 6, we have

$$K\_{di}^{a}≈ \frac{F\_{1}^{'}K\_{d1}^{a}}{F\_{i}^{'}} \left[11\right], K\_{di}^{b}≈ \frac{F\_{1}^{'}K\_{d1}^{b}}{F\_{i}^{'}} \left[12\right]$$

Combining Equations 11 and 12, we obtain

$$K\_{di}^{b}= \frac{K\_{d1}^{b}}{K\_{d1}^{a}}K\_{di}^{a} [13]$$

Take the logarithm of both sides, then we obtain

$$log\_{2}K\_{di}^{b}=log\_{2}K\_{di}^{a}+log\_{2}\frac{K\_{d1}^{b}}{K\_{d1}^{a}} [14]$$

Equation 14 implies that in the graph of $R\_{i}$ (X axis) versus binding energy (Y axis), the Y-axis values of all the data points will change the same amount ($log\_{2}\frac{K\_{d1}^{b}}{K\_{d1}^{a}}$) with the X-axis values remaining unchanged when $K\_{d1}^{a}$ and $K\_{d1}^{b}$ are used. This implies that the residual sum of squares (RSS) of the linear fitting model does not change when $K\_{d1}^{a}$ and $K\_{d1}^{b}$ are used, and the RSS of the constant fitting model does not change either. According to the following Equation 15 computing the F statistic, the resultant F-test value does not change when the two different values $K\_{d1}^{a}$ and $K\_{d1}^{b}$ are used.

$$F= \frac{\left(\frac{RSS\_{1}-RSS\_{2}}{p\_{2}-p\_{1}}\right)}{\left(\frac{RSS\_{2}}{n-p\_{2}}\right)} [15]$$

where $RSS\_{1}$ is the RSS of the linear model, $RSS\_{2}$ is the RSS of the constant model, $n$ is the number of data points (i.e. the number of binding sites in the iPWM), $p\_{2}$ is the freedom of the linear model (i.e. 2), $p\_{1}$ is the freedom of the constant model (i.e. 1).

Then we applied Equation 1 to compute the binding energy of each site.

**2.3) An example slope analysis**

For primary/cofactor motifs, the linear regression model between *Ri* values and binding energy are expected to have slopes well below 0 which is the expected slope for noise motifs.

The following figure shows the probability density distributions of the slopes for 85 primary motifs, 85 cofactor motifs and 85 noise motifs that were randomly selected. The slopes of all primary motifs are smaller than -1.1 except that one is between -1 and -1.1, which also holds true for cofactor motifs; whereas the slopes of all cofactor motifs are greater than -1 except that one is between -1 and -1.1. This implies that the slope of the linear regression model between *log2Kd* and *Ri* can be used to distinguish primary/cofactor motifs from noise motifs.



**References:**

1. Bi,C. and Rogan,P.K. (2004) Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucleic Acids Res.*, **32**, 4979–4991.

2. Levine,H.A. and Nilsen-Hamilton,M. A mathematical analysis of SELEX. *Comput. Biol. Chem.*, **31**, 11–35.

3. Schneider,T.D. (1997) Information content of individual genetic sequences. *J. Theor. Biol.*, **189**, 427–441.

4. Sokolowski,T.R., Walczak,A.M., Bialek,W. and Tkačik,G. (2016) Extending the dynamic range of transcription factor action by translational regulation. *Phys. Rev. E*, **93**, 22404.