Determining Thresholds for Binding Site Sequence Models Using Information Theory

Chengpeng Bi[1,2] Peter K. Rogan[1,2]

[1]Laboratory of Human Molecular Genetics, Children's Mercy Hospital and Clinics, 2401 Gillham Road, MO 64108
[2]School of Computing and Engineering, University of Missouri – Kansas City

Abstract

Models of nucleic acid binding sites based on information theory accurately measure affinities of interactions with regulatory proteins. We present a new algorithm for delineating the constraints on these models. The optimal sequence pattern is found by minimizing entropy of multiple local bipartite or single block alignments. A scanning method was developed to define the extent of sequence conservation for these models (site widths) and to estimate binding site strength cutoffs via Monte Carlo simulation. We applied our methods to the bipartite binding site sequences of *E. coli* cyclic AMP receptor protein and single-block models of splicing regulatory proteins that recognize exonic splicing enhancers.

Keywords: bipartite modeling, information theory, transcription factor, gene regulation, regulation of mRNA splicing.

1. Introduction

To delineate genetic networks, we build experimentally-validated models for each family of binding sites recognized by the same nuclear regulatory factor (TF; [1]). Among TF binding sites, the bipartite motifs are widely recognized as important for gene regulation in both prokaryotes and eukaryotes [2]. A bipartite site consists of two adjacent blocks separated by variable length nucleotide spacer. A bipartite binding site model represented by position weight matrices: M_L for left motif (L) and M_R for right motif (R), and a gap uncertainty function based on the distance separating them, $\omega(d)$. Let $\mathbf{M} = (M_L, M_R)^T$ and D be the gap range allowed, $D = \{d: d_{min} \le d \le d_{max}\}$ and d is an integer. We use the notation: ($\mathbf{M}$, $\omega(D)$), to describe a bipartite model. We use J_L<d>J_R to express a bipartite motif/pattern where J_L and J_R are widths of left and right-site motifs, respectively. Let p_{mj} be a row j vector in a bipartite matrix $\mathbf{M}$. There are 4 nucleotides X = {A, C, G, T} making a DNA sequence, so the size of $\mathbf{M}$ is $|X|\times(J_L+J_R)$. Let $p_{mj}(x)$, an element of vector p_{mj}, be the probability of the nucleotide x at position j of motif $m \in \{L,R\}$. Given a set of known bipartite binding sequences, it is possible to apply minimum entropy-based multiple local bipartite alignment [2] to build such models, which can then be used to scan the relevant genome. In this paper, methods are introduced to automate detection of optimal motif widths and determine threshold information contents necessary to minimize detection of false positives. The optimum width is found using an initial model that is progressively extended with bipartite search patterns to determine the maximum information increment per nucleotide. To simplify the bipartite pattern scanning procedure, we define three information thresholds for the left and right-half sites and bipartite site. A valid bipartite site has information content above all specified cutoffs. We applied these methods to defining *E. coli* CRP transcription factor binding sites, and to sites bound by splicing regulatory factors, SF2-ASF, SC35 and SRp40.

2. Methods

Objective function

Based on experimentation, a set of DNA sequences ($\boldsymbol{S}$) is known to possess bipartite binding sites (either zero or one site per sequence), however the precise start positions of the motif have not been determined. Our goal is to locate those positions and then build a model $\mathbf{M}$ and derive the gap distribution function, $\omega(D)$ for the motif. To infer the parameters of $\mathbf{M}$, we formulate the following sequence likelihood model which is a product of three probabilities, background (p_0), gap (D) and a bipartite site:

$$p(S \mid M_L, M_R, p_0, A, D) \propto p_0^{N(A^c)} \Omega(D) \prod_{m\in\{R,L\}} \left\{ \prod_{j=1}^{J_m} p_{mj}^{N(A_j)} \right\} \quad (1)$$

where p_0 is the background distribution and we assume a uniform: p_0=(0.25, 0.25, 0.25, 0.25). A is the set of unknown bipartite sites (the multiple local bipartite alignment space), A_j is a subset of sites on location j. $N(A_j)$ is the total count of four nucleotides on position j. A^c is the set of nucleotides in background sequences.

We define the quantity $p^N = \prod_{x \in X} p_j^{n_j(x)}$, $n_j(x)$ is count of nucleotide x at position j, $N = \sum_x n(x)$. D is the set of gap distance (d) for all aligned bipartite sequences. $\Omega(D) = \prod_{i=1}^{|S|} \omega(d_i)$. |S| is the total number of training sequences.

Note that one-block models are built by setting the gap size at zero in each sequence. To estimate the optimal parameters for equation 1, we can maximize the log likelihood. This is difficult to carry out directly, as A comprises the hidden or missing data. The expectation maximization and Gibbs sampling methods have been used previously to determine these parameters [2]. We proposed and implemented a so-called minimum entropy-based method [2].

We define the information content (IC) at position j of motif m as the difference between background and motif entropy: $IC_{mj} = H(p_0) - H(p_{mj})$, a special case of relative entropy or the Kullback-Leibler distance [10]. Note that relative entropy or KL distance is defined as $D(p \| p_0) = \sum_{x \in X} p(x) \log p(x) - \sum_{x \in X} p(x) p_0(x)$, and if we assume p_0 follows uniform distribution, then we have $D(p \| p_0) = H(p_0) - H(p) = \log |X| - H(X)$. Therefore, the higher the information content, the greater the level of motif conservation. Our goal is to find motifs with the highest information contents. Instead of directly maximizing the log likelihood, we greedily search the multiple local alignment space and locate those sites (a_{mj}) having the maximum total information content: $\sum_m \sum_{j=1}^{J_m} IC_{mj}$. The maximum likelihood problem is equivalent to maximum information [5]. The entropy $H(p_{mj})$ depends on the aligned motif location (a_{mj}), so we write it as $H(a_{mj})$. Since $H(p_0) = log_2|X| = 2.0$ (bits), the objective function can be reduced to minimize the bipartite entropy (H) conditioning on a gap length range as:

$$(A^*) = \arg \min_{a_{mj} \in A} \left\{ \sum_m \left(\sum_{j=1}^{J_m} H(a_{mj}) \right) \right\} \quad (2)$$

subject to:

$$H(a_{mj}) = -\sum_{x=A}^{T} \hat{p}_{mj}(x_{a_{mj}}) \log_2 \hat{p}_{mj}(x_{a_{mj}})$$

$$\hat{p}_{mj}(x_{a_{mj}}) = \frac{n(x_{a_{mj}}) + \beta_x}{|S| + \sum_{x \in X} \beta_x}, m \in \{L, R\}$$

$$d_{\min} \le a_R - a_L - 1 \le d_{\max}$$

where $A^* = \{a^*_{mj}\}$ representing the optimal bipartite alignment which corresponds to the optimal bipartite model ($\mathbf{M}$, $\omega(D)$). x_{amj} is the nucleotide x at motif (m) position j on sequence starts at a. β_x is the pseudo-count [2] of the nucleotide x. a_L and a_R are start positions for left and right motifs respectively. Equation 2 is a conditional minimization problem. We used a greedy algorithm to search the multiple bipartite alignment space (A) and derive the optimal bipartite model as detailed in [2].

Individual information contents of bipartite sites

We define the total information content (IC) of a bipartite site as sum of left (L) and right (R) half-site information minus a gap penalty ($g(d)$) [2]. Without the gap penalty, IC reduces to R_{sequence}, the average information, for single block binding sites [4]. By considering both strands of each half-site, there are four possible orientations or bipartite patterns [2]. Given a motif start site (a) and bipartite model, the individual information content of that site, R_i, is a function of $\mathbf{M}$, a and d:

$$R_i(M, a, d) = \sum_m \sum_{j=1}^{J_m} \left[2 - \{ -\log_2(p_{mj}(x_{a_{mj}})) \} - e(n) \right] - g(d) \quad (3)$$

where $e(n)$ is a sample correction [4]. The term [2.0-$\{-\log_2(p_{mj})\}$] is the information weight at position j of motif m. Obviously, the nucleotide position with higher probability at j is assigned a higher weight, such that invariant positions exhibit the maximum R_i value, 2.0 bits. $R_i = 0$ bits when all nucleotides have the probability of the background sequence, ie. 0.25. A negative weight will be assigned for probabilities below 0.25. The gap penalty function is based on the gap frequency distribution ($\omega(D)$) which is derived from the best bipartite alignment (A^*). We define the gap frequency: $\omega(d) = f(d)/|D|$, $f(d) = 1.0 + cos(2\pi(d - c)/B)$, B is a DNA helical repeat (10.4 bases/turn). So this gap function is applied to a short-gapped bipartite site (i.e. $d_{max} \le 10$ bps). c is the dominant or central gap size, $f(c) = 2.0$. The normalized gap penalty is defined by,

$$g(d) = -\log_2(\omega(d)) + \log_2(2.0/|D|) = 1.0 - \log_2 f(d) \quad (4)$$

As shown in equation 4, the function assigns a zero penalty to the dominant gap size.

Refinement of bipartite patterns

Given the initial bipartite binding site pattern: $\Delta^{(0)} = J_L^{(0)} < D > J_R^{(0)}$, J_L and J_R are the left and right motif lengths, respectively, and D is the gap range [d_{min}, d_{max}]. We assume that the bipartite length range is [L_{min}, L_{max}]. With the constraints: $J_L + J_R + d_{min} \le L_{min}$ and $J_L + J_R + d_{max} \le L_{max}$, we can generate a certain number of search patterns (Ψ). Based on the initial search pattern, a bipartite model can be derived with the average total information content, $IC^{(0)}$. Let the t-th search pattern be $\Delta^{(t)} = J_L^{(t)} < D > J_R^{(t)}$ and average information content be $IC^{(t)}$. We define a unit information incremental index (UII) as:

$$UII^{(t)} = \frac{IC^{(t)} - IC^{(0)}}{(J_L^{(t)} + J_R^{(t)}) - (J_L^{(0)} + J_R^{(0)})} \quad (5)$$

The optimal pattern (Δ*) selected exhibits the highest information increment:

$$\Delta^{(*)} = \arg\max_{t \in \Psi} \{UII^{(t)}\} \quad (6)$$

We can think of the initial pattern as the basis (core motifs) of a biological motif. We then extend both half-site motifs and find a new pattern with maximum UII such that $J_L^{(new)} \geq J_L^{(0)}$ and $J_R^{(new)} \geq J_R^{(0)}$. If more than one maximum pattern is found, we take the pattern with the largest extension among ($J_L + J_R$).

Bipartite scanning model

To search for a bipartite instance of a known model in a DNA sequence (s) we built the probabilistic model for a bipartite site in a sequence (equation 7). The basic sequence model assumes that binding sites are embedded in "noisy" background sequence which is assumed to follow uniform multinomial distribution (p_0). The bipartite model consists of two half-site PWM-based models (M_L and M_R) and a gap function $\omega(d)$. If the start position of the bipartite instance is known and indicated by a, and the gap distance is d, then the probability that the sequence is generated given the model parameters is,

$$P(s \mid M_L, M_R, p_0, a, d) \propto \prod_{l \in a^c} p_0(x_l) \times \prod_{m \in \{R,L\}} \left\{ \prod_{j=1}^{J_m} p_{mj}(x_{a_{mj}}) \right\} \times \omega(d) \quad (7)$$

where $x_l \in X$, $p_{mj}(x)$, an element of **M**, is the probability of finding the nucleotide x at motif position j and a^c is the background sequence. Since we assume a constant background, we simply find a site with individual information content above a threshold.

Consider the likelihood of observing a bipartite motif generated from sequence with pure background distribution p_0 versus the corresponding probability a second sequence generated from p_0 with a number of embedded bipartite sites, according to (**M**, $\omega(D)$). Instead of directly computing the bipartite motif probability for a sequence in equation (7), we calculate the individual IC (R_i) for a site separately as described in equation (3), information for background sequences with embedded sequences. We wish to determine the likelihood of finding real sites that are true positives. We determine the IC distributions for different scenarios by generating background sequences, scanning the sequences with the previously developed information models (**M**, $\omega(D)$), calculating the individual IC's for all possible motif sites, and plotting the distribution. We assume that IC values follow a normal distribution with IC>0 bits [as we have previously shown, 6], which can be easily derived from bipartite training dataset. Binding site scans of the background sequences containing the embedded binding sites versus background sequences alone determine IC thresholds that distinguish true bipartite sites from decoys. The positive bipartite IC distribution derived from background sequences is then compared with the known bipartite IC distribution from the samples.

Monte Carlo simulation

We generate a set of background DNA sequences according to the distribution p_0 and calculate information content (R_i) for each of scanned bipartite sites according to equations 3 and 4. Our bipartite model is based on information theory, so the background motif distribution is assumed as uniform multinomial, p_0. The procedure for generating the sequences (dataset) was the same as previously described [2]. The frequency (probability) distribution of non-site information content will be derived from the simulated datasets. Based upon the fact that a binding site is considered as a putative site if its IC exceeds zero bits, we want to derive a conditional distribution $P(\ S \mid IC>0,\ \mathbf{M})$ for both random (by chance) and embedded (known) bipartite sequences. The planted site distribution is derived from known sequences and assumed a normal distribution. The distributions (half-sites and total individual IC) for random sequences are derived by simulation. Binding site threshold values (C_b for bipartite IC, C_l and Cr for left and right-half site IC's) are determined by minimizing Bayes error rates (Type I and II) [11]. True bipartite sites are determined subject to IC(bipartite) > C_b and IC(left) > C_l and IC(right) > C_r.

3. Results

CRP binding sites

A bipartite model was developed for a set of binding sites consisting of 18 sequences recognized by the dimeric cyclic AMP receptor protein (CRP). Each sequence is 105 base pairs long and each contains at least one bipartite site that has been experimentally identified via footprinting. The initial pattern was initially defined as a short-gapped bipartite motif: 5<1,9>6 and $11 < J_L + J_R + d < 23$ (bps). The left and right motif widths were extended up to 10 bps, respectively. In total, we applied multiple bipartite local alignment based on minimum entropy to 30 search patterns. Bipartite models for each pattern were derived and UII's were calculated for each (Fig. 1). The optimal search patterns are 5<1,9>8 and 6<1,9>6 as seen the maximum peaks (UII values are 0.44 and 0.45 bits per base, respectively). We take 5<1,9>8 as the optimal bipartite pattern.

Treating the CRP binding site as an imperfect direct repeat, we extended the optimal pattern 5<1,9>8 to 8<1,9>8 and recomputed the UII value at 0.238 bits per base. Fig. 2 shows the optimal bipartite logo of CRP binding sites [2,3], containing an imperfect palindromic sequence, which is in agreement with our previous report [2].

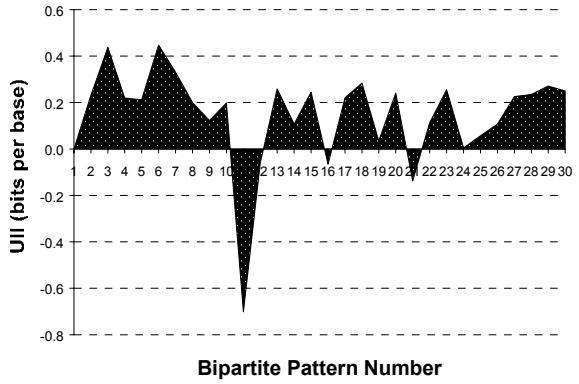

Fig. 1 Unit information increment index (UII) for 30 different bipartite search patterns. 500 cycles were run for each [3].

Estimation of IC thresholds via simulation

Although information theory predicts that sites exceeding zero bits should be recognized, experimentally studies do not always detect weak sites with low ICs. The minimum IC for the CRP 8<6>8 model was estimated by simulation.

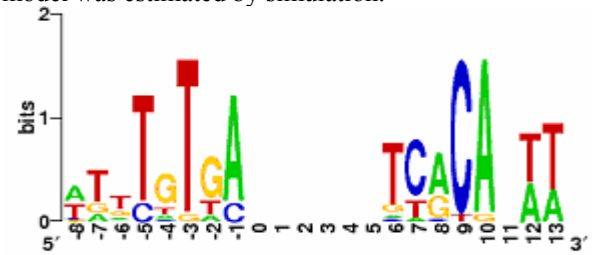

Fig. 2 Bipartite sequence logo [2] for CRP bipartite binding model 8<6>8, central gap $c = 6$.

The bipartite model was used to scan 10^6 simulated sequences (each 100 bp long) and to estimate the site frequencies for left and right-half sites and bipartite sites conditioning on IC> 0.0 bits. To determine the IC thresholds, we plot the random sites together with the known site distributions and graph the IC threshold values in each case. The derived cutoffs are: $C_l = 3.8$, $C_r = 4.2$ and $C_b = 8.6$ bits. The Bayes error rates for bipartite IC are lower (0.015) than half-site IC (0.15). The IC distribution for the bipartite site is indicated in Fig. 3 where the dotted line demarcates the Bayesian decision boundary giving the lowest probability of error [11]. Based on these thresholds, we are scanning the *E. coli* genome with the CRP bipartite model to search for putative binding sites upstream of co-regulated genes (not shown).

Application to regulatory splicing binding sites

We estimated the site widths and IC thresholds of experimentally-validated binding sites recognized by proteins that regulate mRNA splicing: SC35, SRp40 and SF2-ASF [7,8]. The model for each site was initially set to a width of 5 bp and then extended. Fig. 4 shows the results of one-block motif UII curves. Three binding motifs have peaks at 6 bps. TheSF2-ASF curve is uni-modal, and SC35 and SRp40 curves are multi-modal. The SC35 binding motif has second peak at 8 bps. A very weak peak for SRp40 occurs at 8 bps. Given the UII cutoff of 0.05 bits per base, the optimal widths are 7, 6 and 8 bps for SF2-ASF, SRp40 and SC35 protein binding sites, respectively (sequence logos presented in Table 1). For a multi-modal UII curve, here we set a cutoff UII (e.g. 0.05) and take the longest motif (e.g. 8 bps) as the optima.

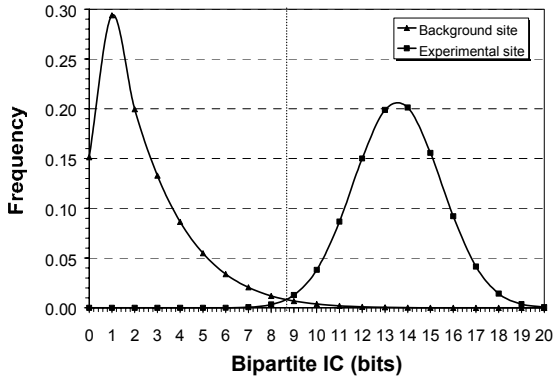

Fig. 3 Plots of IC distributions for background vs experimentally-determined bipartite CRP sites.

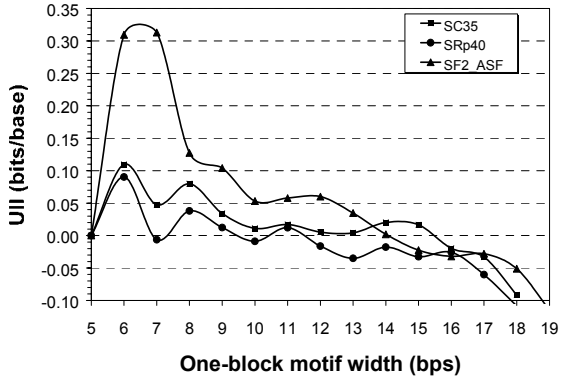

Fig. 4 Plots of motif width vs. UII for splicing regulatory proteins: SC35, SF2-ASF and SRp40.

Table 1 shows the optimal widths (w) for single stranded mRNA splicing regulatory protein binding sites, IC thresholds (δ), average information contents (IC), and sequence logos. It also shows that 7-mer motif for the optimal width of SF2-ASF protein binding sites is preferred as more information gained and conserved. Very little information is gained for

SRp40 8-mer motif compared to the 6-mer and the models are statistically indistinguishable [4].

Table 1. Optimal widths (w) and IC thresholds (δ) of splicing regulatory protein binding sites

Factor	w	δ	IC	Sequence logo
SC35	6	2.5	4.39	
	8	2.7	4.53	
SF2	7	3.6	5.77	
SRp40	6	2.8	4.54	

4. Discussion

In this paper we gave a systemic account of bipartite modeling and scanning problems. Given an initial search pattern, we applied greedy algorithm to do minimum entropy-based alignment and build the bipartite model. We then extend the initial pattern to the point where the incremental information is maximized. Three information cutoffs for a bipartite site are determined via Monte Carlo simulation.

To implement effective threshold determination of bipartite modeling, we suggest a reasonable initial pattern of core motif widths consistent with the biological data, since the full length of bipartite motif is usually unknown and short repeats are frequently present in regulatory regions of genomes [2]. The final adjustment relies on the fact that the most common nucleotides in a bipartite (or single block) logo are related to inferred consensus patterns. The models often, however, depict subtle variation at conserved positions. The threshold algorithms presented here can potentially reveal alternative binding configurations and delineate weak binding sites not evident from consensus sequences. Despite the presumed increased sensitivity for recognizing true binding sites, it is unlikely that *de novo* automation of motif discovery would be feasible without experimental verification.

If the calculated UII value falls below the UII cutoff, we treat the additional nucleotide information as noise and thus ignore these positions; otherwise, they are considered signals and retained in the weight matrix.

We set three cutoffs for each scanned bipartite site in order to reduce the false positive rate while scanning a genome. The disadvantage of doing so is that it potentially misses some weak binding sites as the thresholds were derived based on minimizing Bayes error-rate.

Significant differences between models of binding sites recognized by the same factor are sometimes evident, when the IC is the only quantitative criterion to define sites (Table 1). Minimizing uncertainty can result in multiple equivalent outcomes, such as that seen for SC35. Experimental mutagenesis and binding studies are the only means of unequivocally distinguishing the correct model.

The optimized mRNA splicing regulatory protein models derived in this study can be used to predict human mutations at these sites at https://splice.cmh.edu; [9].

5. Acknowledgements

Support from the Katharine B. Richardson Foundation and ES 10855 is gratefully acknowledged.

6. References

[1] C. Vyhlidal, P. Rogan, J. Leeder "Development and Refinement of Pregnane X Receptor DNA Binding Site Model Using Information Theory", *J. Biol. Chem*, 279: 46779 – 46786, 2004.

[2] C. Bi and P. Rogan "Bipartite Pattern Discovery by Entropy Minimization-Based Multiple Local Alignment," *Nucl. Acids Res.,* 32, 4979-4991, 2004.

[3] Bipad server: http://bipad.cmh.edu.

[4] T. Schneider, G. Stormo, L. Gold, and A. Ehrenfeucht "Information content of binding sites on nucleotide sequences," *J. Mol. Biol*., 188: 415-431, 1986.

[5] T. Bailey "Likelihood vs. Information in Aligning Biopolymer Sequences", USCD Technique Report CS93-318, 1993.

[6] P. Rogan, S. Svojanovsky and J. Leeder "Information theory-based analysis of *CYP2C19*, *CYP2D6* and *CYP3A5* splicing mutations," *Pharmacogenetics* 13: 207-218, 2003.

[7] H-X. Liu, S. Chew, L. Cartegni, M. Zhang and A. Krainer "Exonic splicing enhancer motifs recognized by human SC35 under splicing conditions," *Mol. Cell Biol.* 20: 1063-1071, 2000.

[8] H-X. Liu, M. Zhang and A. Krainer "Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins," *Genes & Devel.* 12: 1998-2012, 1998.

[9] V. Nalla and P. Rogan "Automated splice site mutation analysis," *Hum Mut*. 25:334-342, 2005.

[10] T. Cover and J. Thomas "Elements of Information Theory", Wiley & Sons, Inc, 1991.

[11] R. Duda, P. Hart and D. Stork "Pattern Classification", Wiley & Sons, Inc, 2001.