Brown University

Finding Motifs Using Gibbs Sampling

Anirudh Narsipur

December 17, 2021

1 Introduction

Finding motifs for transcription factors is important in the study of transcription factors. However, finding correct motifs efficiently is not trivial.

Here I implement Gibbs Sampling, a heavily studied algorithm for finding motifs. I apply this algorithm to find motifs for transcription factors in Mycobacterium tuberculosis (MTB) and discuss my results.

2 Background

Within each cell, DNA gets transcripted into RNA which is then translated into protein. The amount of expression of a protein and it's timing is strictly controlled. This ensures that proteins are expressed at the correct stage in the cell life cycle or not expressed at all. Transcription factors (TF) are an important part of this regulatory framework.

Transcription factors are proteins that bind slightly upstream of the genes they regulate. They either promote (activators) or block (repressors) gene transcription by RNA polymerase. Transcription factors are usually highly specific and bind to specific sequences known as DNA binding domain (DBD).DBDs contain a sequence motif that is a conserved DNA sequence. Transcription factors have one or more motifs associated with them and they only bind at locations where the motif is present.

Thus it important to identify motifs associated with each transcription factor in order to study and classify them. Through experimental techniques, such as ChIP-Seq

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}.$$

Figure 1: An example of a Position Weight Matrix (PWM) for a DNA sequence of length 9.

the sequences around the binding sites of transcription factor can be identified. The computational task is to then identify the motif given the sequence data.

In this project I use Gibbs Sampling to identify motifs and then apply it to Chip-Seq data to identify the motifs in Mycobacterium tuberculosis. Gibbs Sampling is a probabilistic algorithm that iteratively works towards optimal motifs. From the algorithm we can extract motifs as position weight matrices (PWM).

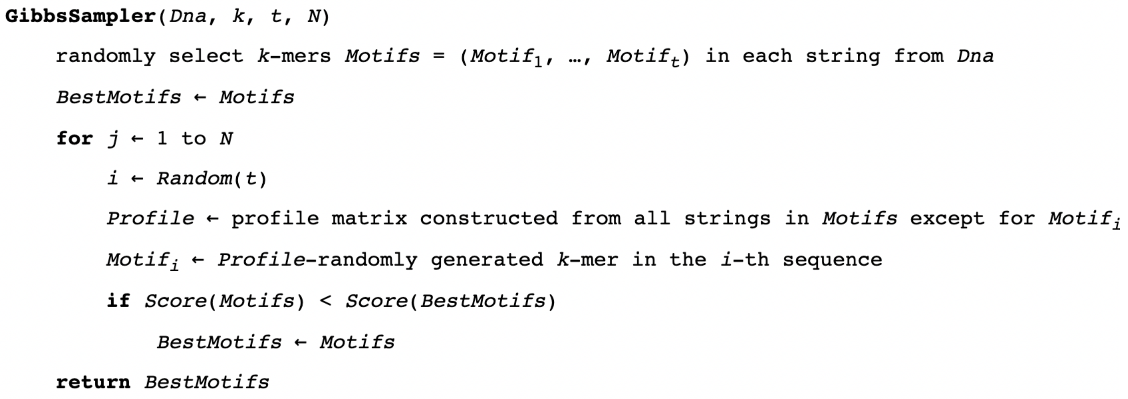
Position weight matrices as given above represent a motif by the probability of a each nucleotide being present at a given position.

3 Methods

To rigorously frame our problem of find motifs:

Given n DNA sequences s of length L, and a target motif length k, find a $k-mer$ in each string s_i such that distance between s_i and s_j is minimized. $\forall i,j \in \{1,n\}$

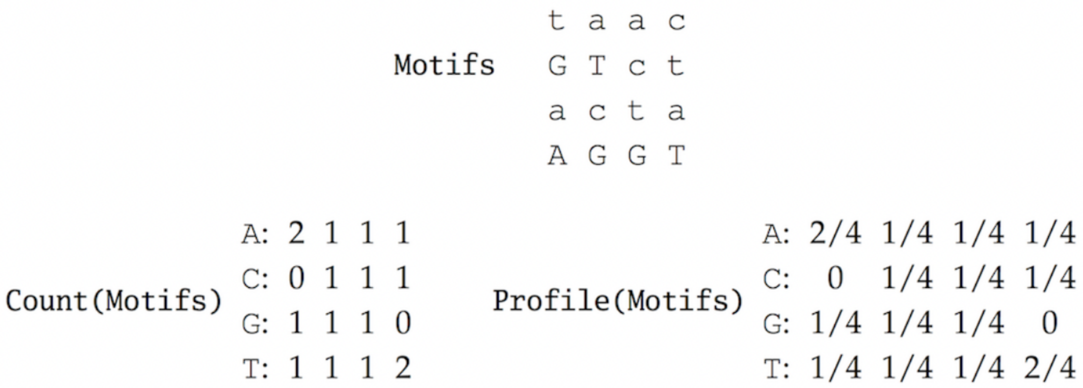
We then use Gibbs Sampling to solve it. Gibbs Sampling proceeds iteratively in the following manner:



```
GibbsSampler(Dna, k, t, N)
    randomly select k-mers Motifs = (Motif_1, …, Motif_t) in each string from Dna
    BestMotifs ← Motifs
    for j ← 1 to N
        i ← Random(t)
        Profile ← profile matrix constructed from all strings in Motifs except for Motif_i
        Motif_i ← Profile-randomly generated k-mer in the i-th sequence
        if Score(Motifs) < Score(BestMotifs)
            BestMotifs ← Motifs
    return BestMotifs
```

Fig 2. Pseudocode for Gibbs Sampling.

- In each string s_i, pick a random $k-mer$ x_i.
- Randomly select a string s_j from the set of strings s.
- Construct a profile matrix P from the strings $s - \{s_j\}$.



Motifs

t	a	a	c
G	T	c	t
a	c	t	a
A	G	G	T

Count(Motifs)

A:	2	1	1	1
C:	0	1	1	1
G:	1	1	1	0
T:	1	1	1	2

Profile(Motifs)

A:	2/4	1/4	1/4	1/4
C:	0	1/4	1/4	1/4
G:	1/4	1/4	1/4	0
T:	1/4	1/4	1/4	2/4

Fig 3. Construction of the profile matrix.

The profile matrix P is constructed by taking the count of each nucleotide in each position of the strings and then dividing by the total number of sequences considered $(n-1)$. However, observe that some values are zero. This leads to problems in scoring k-mer as given below. Having a value of zero in the first column of the profile matrix for C indicates that probability of motif starting with C is mathematically zero.This is biologically not true as even if a C does not occur at the first nucleotide in the best motif we find later that does not mean the chance of it occurring is zero.
To correct for this we add a ”pseudocount” to the profile matrix. While more sophisticated methods exist, here we use a simplistic method based on Laplace's rule of succession, that is for each value in the profile matrix m we do $\frac{m+1}{(n-1)+4}$.

- Score all $k-mers$ in s_j against the profile matrix P. $k-mer$ are score by simply taking the product of the probability of each nucleotide at each position in the k-mer as given in the profile matrix. To put it more formally,

$$Score(x) = Pr(x|P) = \prod_{i=1}^{L} e_i(x_i)$$

  where $e_i(x_i)$ is the probability of observing the $i'th$ in x given M.

- Normalize the resulting probabilities to get a probability distribution D. For a given sequence we have $w = L - k + 1$ possible k-mers. If our scores are $c_1 ... c_w$,then $D = \{\frac{c_i}{\sum_{j=1}^{w} c_j}, \forall i \in [1, w]\}$. To put it simply, we divide each probability by the sum of all probabilities.

- Sample a new motif x_j from the probability distribution D.

- Repeat the above process until convergence or for a set number of iterations.

3.1 Background Model

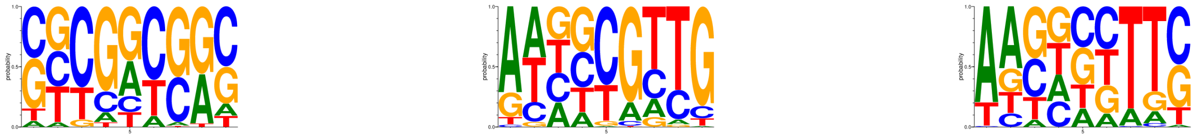
The above model fails to consider several biological facts:

- Nucleotides are not independent. The probability of observing a nucleotide x_i in a given position i is not independent of the sequence preceding it.
- Nucleotides are usually not equally distributed, and many organisms tend to be either GC dominant or AT dominant. These percentages can significantly vary in different parts of the genome.
- Repeated sequences can be a confounding factor. For example, *GGG* could be a common repetition across sequences but not part of a motif.

A powerful way to overcome these limitations is to use a background model B.A background model B_i has order i indicating that the probability of observing a nucleotide depends on i previous nucleotides.

4 Results

In order to validate the implementation of the methods described above Chip-Seq derived data from Minch et al. (2015) was used. The dataset contained approximately 16,000 unique peaks for 156 regulators. In order to avoid issues with overtly small data sizes I got results only from regulators with more than 50 peaks. I ran the Gibbs Sampling algorithm with no background model, a zero order background and a first order background model. All motifs generated are of length 9 and the algorithm was run for 3000 iterations.



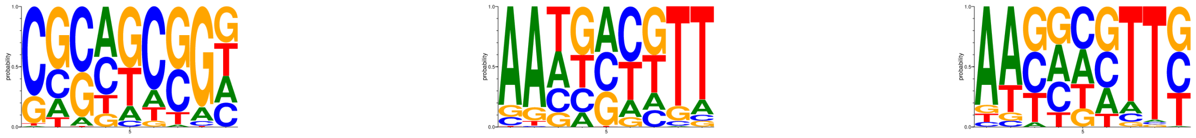
(a) No background model

(b) Zero background

(c) First order background

Figure 2: Motifs Visualized For Rv002c

Visualized above are motifs generated for Rv002c and Rv2009 which 218 and 67 samples respectively. Immediately we can see a dramatic difference in using a background



(a) No background model

(b) Zero background

(c) First order background

Figure 3: Motifs Visualized For Rv2009

model. Since Mtb is GC rich without a background model we pick out GC repeats. Using a background model we pick out much better motifs that are not simple repeats.

Interestingly, the information content between different background models does not differ.A first order background model leads to slightly lower information content. It is impossible to draw strong conclusions from the above results given the small sample size (51 regulatory factors). Further information content clearly does not tell the whole story here as can be seen from the sequence logos.

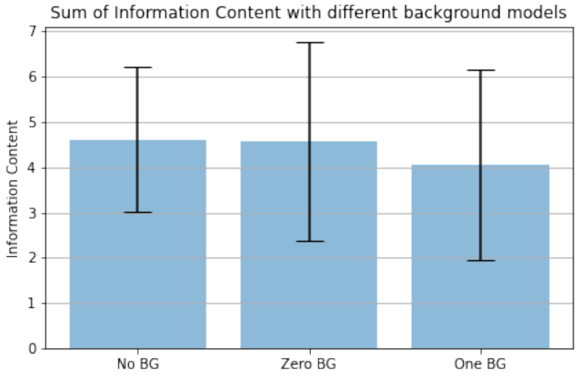


Figure 4: Information Content Between Background Models

5 Acknowledgements

This project is a study of well established methods and as such would not possible without existing literature.A special debt is owed to Bioinformatics Algorithms by Compeau and Pevzner from where all figures are taken from. All data used in this project is drawn from Minch et al. (2015).

6 References

- Phillip Compeau, Pavel Pevzner (2011). BIOINFORMATICS ALGORITHMS.Link
- Minch, K., Rustad, T., Peterson, E. et al. The DNA-binding network of Mycobacterium tuberculosi s. Nat Commun 6, 5829 (2015). https://doi.org/10.1038/ncomms6829
- Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouzé P, Moreau Y. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics. 2001 Dec;17(12):1113-22. doi: 10.1093/bioinformatics/17.12.1113. PMID: 11751219.
- A Gibbs Sampling Method to Detect Overrepresented Motifs in the Upstream Regions of Coexpressed Genes Gert Thijs, Kathleen Marchal, Magali Lescot, Stephane

Rombauts, Bart De Moor, Pierre Rouzé, and Yves Moreau Journal of Computational Biology 2002 9:2, 447-464

- Seshasayee AS, Sivaraman K, Luscombe NM. An overview of prokaryotic transcription factors : a summary of function and occurrence in bacterial genomes. Subcell Biochem. 2011;52:7-23. doi: 10.1007/978-90-481-9069-0_2.

  PMID: 21557077.

- Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 1993 Oct 8;262(5131):208-14. doi: 10.1126/science.8211139. PMID: 8211139.

- George Casella , Edward I. George (1992) Explaining the Gibbs Sampler, The American Statistician, 46:3, 167-174, DOI: 10.1080/00031305.1992.10475878