Obtain total affinity and occupancies for binding site matrices on a given sequence

Elena Grassi
Department of Molecular Biotechnologies and Health Sciences
MBC, University of Turin, Italy
grassi.e@gmail.com

MatrixRider version 1.4.0 (Last revision 2015-02-10)

Contents

Abstract

Transcription factors regulate gene expression by binding regulatory DNA: understanding the rules governing such binding is an essential step in describing the network of regulatory interactions, and its pathological alterations.

This package implements a method that represents an alternative to classical single site analysis by summing all the single subsequence affinity contributions of a whole sequence, representing an approach that is more in line with the thermodynamic nature of the TF-DNA binding.

1 Introduction

The first step in understanding transcriptional regulation consists in predicting the DNA sequences to which a TF is able to bind, so as to identify its targets. Most TFs bind sequences that are relatively short and degenerate, making this prediction quite challenging. The degeneracy of the binding sites is reflected in the use of a Positional Weight Matrix (PWM) to describe the binding preferences of a TF. A PWM specifies the frequency distribution of the 4 nucleotides in each position of a binding site, and is typically used to assign a score to each DNA sequence. Roughly speaking the score expresses the degree of similarity between the observed sequence and the PWM. A sequence is then predicted to be a transcription factor binding site (TFBS) if it scores above a given cutoff.

The introduction of a cutoff is unsatisfactory not only because it introduces an arbitrary parameter, but also and especially because recent detailed investigations of transcription factor binding have shown it to be a thermodynamic process in which transient binding to low-affinity sequences plays an important role. In this view the concept itself of a binary distinction between binding and non-binding sites comes into question: it becomes more appropriate to consider the total binding affinity (TBA) of a sequence taking contributions from both high- and low-affinity sites [**?**].

This approach was indeed pioneered and applied to transcriptional regulation in yeast by the Bussemaker lab [**?**, **?**]. Recently we used total binding affinity profiles to study the evolution of cis-regulatory regions in humans [**?**] and decided to include our C code used to calculate affinity in a small (but well integrated with Bioconductor TF binding sites resources) package.

We have added the possibility to sum only the affinities larger than a given cutoff instead that all of them to compare the predictive power, regarding real binding events, of both approaches. We refer to "total affinity" when no cutoff is used and to "occupancy" otherwise.

2 Looking for binding potential for a single TF on a sequence

The most straightforward way to use our package is to obtain the binding preferences information for a given TF using *JASPAR2014* and *TFBSTools* and then use the `getSeqOccupancy` with three arguments: a *DNAString* with the sequence of interest, the *PFMMatrix* and a numerical cutoff parameter.

```
> library(MatrixRider)
> library(JASPAR2014)
> library(TFBSTools)
> library(Biostrings)
> pfm <- getMatrixByID(JASPAR2014,"MA0004.1")
> ## The following sequence has a single perfect match
> ## thus it gives the same results with all cutoff values.
> sequence <- DNAString("CACGTG")
> getSeqOccupancy(sequence, pfm, 0.1)
```

```
[1] 1470.946
```

```
> getSeqOccupancy(sequence, pfm, 1)
```

```
[1] 1470.946
```

The *PFMMatrix* counts and background information are used to obtain likelihood ratios for all the possible nucleotides in a given sequence. A pseudocount of one is added to the counts that are equal to zero. The cutoff parameter should be comprised between 0 and 1: 1 means summing up only affinities corresponding to the perfect match for the given matrix (i.e. for MA0004.1 the sequence "CACGTG"[1]). 0 corresponds to the so called "total affinity": every affinity value is summed. All the other values represents trade-offs between these two extremes. For more details on the performed calculation see **??**.

3 Working with multiple matrixes

Another possible approach is to use as argument a *PFMMatrixList*: in this case the return value is not a single number but a numeric vector with all the obtained affinites on the given *DNAString* for the given matrixes. It will retain the names of the *PFMMatrixList*.

```
> pfm2 <- getMatrixByID(JASPAR2014,"MA0005.1")
> pfms <- PFMatrixList(pfm, pfm2)
> names(pfms) <- c(name(pfm), name(pfm2))
> ## This calculates total affinity for both the PFMatrixes.
> getSeqOccupancy(sequence, pfms, 0)
```

```
    Arnt       AG
1470.946    0.000
```

In the examples of the package you can find a simple R script that calculates affinities for all the Vertebrates matrixes found in *JASPAR2014* for a given multifasta file. It is also possible to use manually made (i.e. derived from other databases different than Jaspar) matrixes: one simply needs to build a *PFMMatrix* object with the desired counts (need to be integer values) and the background frequencies.

[1] the perfect match of a given matrix could change with different background distribution values of nucleotides

4 Appendix A

Total affinity is defined as in [**?**]: a_{rw} of a PWM w for a sequence r is given by:

$$a_{rw} = \log \sum_{i=1}^{L-l} \max \left(\prod_{j=1}^{l} \frac{P(w_j, r_{i+j})}{P(b, r_{i+j})}, \prod_{j=1}^{l} \frac{P(w_{l-j+1}, r'_{i+j})}{P(b, r_{i+j})} \right)$$

where l is the length of the PWM w, L is the length of the sequence r, r_i is the nucleotide at the position i of the sequence r on the plus strand, r'_i is the nucleotide in the same position but on the other strand, $P(w_j, r_i)$ is the probability to observe the given nucleotide r_i at the position j of the PWM w and $P(b, r_i)$ is the background probability to observe the same nucleotide r_i.

To apply a cutoff similar to the one used when defining single binding events that relies on the maximum possible score for a PWM we had to express the fractional cutoff, that is normally calculated on the log likelihood of a sequence of length l, referring only to the $P(w_j, r_i)$ ratios.

Assuming a fractional cutoff c we wanted to sum only the scores for the positions on sequences that correspond to log likelihoods bigger than or equal to $c \times \sum_{j=1}^{l} \log(\frac{P(w_{jPWM}, r_j)}{P(b, r_j)})$, where w_{jPWM} is the nucleotide with the higher ratio between the binding model and background probabilities in the PWM at position j.
This corresponds to

$$\max \left(\prod_{j=1}^{l} \frac{P(w_j, r_{i+j})}{P(b, r_{i+j})}, \prod_{j=1}^{l} \frac{P(w_{l-j+1}, r'_{i+j})}{P(b, r_{i+j})} \right) \geq \prod_{j=1}^{l} \left(\frac{P(w_{jPWM}, r_{i+j})}{P(b, r_{i+j})} \right)^c$$

assuming that we are working on the subsequence of r that begins at position i. We will refer to this disequality as $PWM_c(c, w, r, i)$ from now on.

Thus we define the total occupancy t_{rwc} of a PWM w for a sequence r and cutoff c with $0 \leq c \leq 1$ as:

$$t_{rwc} = \sum_{i=1}^{L-l} \max \left(\prod_{j=1}^{l} \frac{P(w_j, r_{i+j})}{P(b, r_{i+j})}, \prod_{j=1}^{l} \frac{P(w_{l-j+1}, r'_{i+j})}{P(b, r_{i+j})} \right) \times \phi(c, w, r, i)$$

with the ϕ function defined as:

$$\phi(c, w, r, i) = \begin{cases} 1 & \text{if } c = 0 \text{ or } PWM_c(c, w, r, i) \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

This definition makes the logarithm of the total occupancy with $c = 0$ identical to the total binding affinity, as is intuitively expected.