

Selection of Putative *Cis*-regulatory Motifs Through Regional and Global Conservation

Youlian Pan¹, Brandon Smith², Hung Fang², Fazel A. Famili¹, Marianna Sikorska² and Roy Walker²

¹Institute for Information Technology and ²Institute for Biological Sciences,
National Research Council Canada, 1200 Montreal Road, Ottawa, Ontario, Canada K1A 0R6
{firstName.lastName}@nrc-cnrc.gc.ca

Abstract

Cis-regulatory motifs are often overrepresented in promoters and may exhibit frequency biases in sub-promoter regions (SPRs). Many probabilistic algorithms have been used to predict such motifs, but they tend to generate many false positives. We devised a novel algorithm, MotifFilter, that computes Representation Indices (RIs) for putative motifs. MotifFilter's RI is a ratio of the actual over expected frequency of a motif in promoters, SPRs or random genomic DNA that takes into account of the nucleotide probability distributions in these regions. This approach was applied to a genome-wide survey of putative cAMP-response elements (CREs) for motifs generated by a profile hidden Markov model. Twenty of 144 putative CRE motifs found in the survey were retained by the MotifFilter.

1. Introduction

Discovery and characterization of *cis*-regulatory sequence motifs in a completed genome is key to the understanding of gene relationships. These motifs can be characterized using probabilistic models, such as profile hidden Markov models (PHMMs). In PHMM, a position-specific conservative profile of residuals is generated from a family of functionally and structurally related protein or DNA sequences using a well-formulated probabilistic model [1]. The probabilistic model is then used to detect a sequence element that is similar to the family. One drawback of such an algorithm is that the number of false positive predictions is often extraordinarily high. We report here a new algorithm that uses frequency distribution characteristics to select for putative motifs that are more likely to be functional.

2. The algorithm

The algorithm consists of two components: calculation of representation indices (RIs) for each

motif in genomic sequences, promoters and SPRs, and filtering based on these parameters (Fig. 1).

2.1. Representation index

Probability distributions $[p(x) \{x = A, T, C, G\}]$ of nucleotides are calculated based on their frequencies in a given set of sequences. The probability of a given motif $[p(M)]$ is derived from these nucleotide probabilities:

$$p(M) = \prod p(x_i) \quad \{i = 1, 2, \dots, m\} \quad (1)$$

where M is a motif, x_i is the nucleotide at i , and m is the length of the motif. Statistical expectations of the motif in the entire sequence dataset (SE) and in an SPR (SE_j) are calculated based on the number of nucleotides that a scan algorithm has covered:

$$SE = p(M) \times (n - m) \times s \quad (2)$$

$$SE_j = p(M)_j \times b \times s \quad (3)$$

where n is the length of each sequence in the dataset (we assume all sequences in the dataset have the same length); s is the number of sequences; j is the SPR; and b is the SPR length. The representation indices of a given motif M in the entire sequence data [$RI(M)$] and in an SPR [$RI_j(M)$] are:

$$RI(M) = \frac{F}{SE} \quad (4)$$

$$RI_j(M) = \frac{F_j}{SE_j} \quad (5)$$

where F and F_j are the frequencies of the given motif in the entire dataset and in an SPR respectively.

2.2. Motif filter

Details of the MotifFilter algorithm are depicted in Fig. 1.

```

MotifFilter( ArrayOfMotifs )
  Generate base probabilities for random genomic sequences
  Generate base probabilities for promoter sequences
  Generate base probabilities for SPRs
  For each motif instance
    Calculate expected motif instances for each dataset (SE)
    Find actual motif instances in each dataset (F)
    Calculate representation index (R) for
      Random genomic sequences →  $RI_g$ 
      All known promoters →  $RI_p$ 
    Get ratio (R) of representation indices ( $RI_p/RI_g$ )
    Calculate expected motif instances in each SPR (SE)
    Find actual motif instances in respective SPRs (F)
    If  $R \geq T_R$ 
      Selected → M1 group
      For each M1 member
        Calculate SPR-specific  $RI_p(\text{SPR})$ 
        For  $j = k, k+1, k+2, \dots, k+y$  { $k=1, 2, \dots, c-y; y < c/2$ }
          If all  $RI_p(\text{SPR}_i \in \mathbf{M}) \gg RI_p(\text{SPR}_i \notin \mathbf{M})$  { $i \neq j$ }
            Selected → M2 group
          Else discard
        Else discard
      Report M2 group
    End
  End

```

Figure 1. The MotifFilter algorithm. T_R is a threshold of R . c is the number of SPRs. \mathbf{M} is a group of consecutive SPRs.

3. Implementation and application

MotifFilter was implemented in the BioMiner (http://iit-iti.nrc-cnrc.gc.ca/projects-projets/biomine_e.html), a suite of tools for data mining in genomics. An application of MotifFilter to a genome-wide survey of CRE motifs is described below.

3.1. Sequences and motifs

Putative promoter regions in this study were defined as the 1000bp upstream of the annotated transcription start site (TSS) of RefSeq (<http://www.ncbi.nlm.nih.gov/RefSeq/>). Human promoter regions were obtained from the UCSC Genome Bioinformatics site (genome version hg16; <http://genome.ucsc.edu/>). A genomic sequence set was generated by randomly selecting $10,000 \times 1000$ bp regions from the entire human genome. One hundred fourteen putative motifs were generated by a PHMM built from the CREB_01 positional weight matrix of TransFac (version 7.2) [2].

3.2. Filter parameters and results

The SPR length (b) was set to 50bp and the R threshold (T_R) was 1.5. The first 300bp upstream of the TSS was used as the consecutive SPR region (\mathbf{M}). Motifs with an $RI_p(300)/RI_p(700) \geq 2.0$ are shown in Table 1. Of the 20 selected motifs, 14 were confirmed either by another TransFac consensus for CREs or by the literature. Figure 2 shows that regional motif conservation may be enhanced by incorporating regional nucleotide distributions.

Table 1. Filtering results of putative CREB_01 motifs. The motif in *italics* is used in Figure 2.

Motif	R	$RI_p(300)$	Confirmed by TransFac consensus or literature
		$RI_p(700)$	
TGACGTCA	10.50	9.16	CREB_01
TGACGTAA	3.48	5.73	CREB_01
TGACGTAT	2.46	2.97	CREB_02
TGACGTCG	10.62	2.92	CREB_02
TGACGTAG	2.64	3.65	CREB_02
TGACGCCA	2.54	2.60	CREB_02
TTACGTCA	3.48	5.73	
TGACGCAA	4.28	3.87	CREB_02
<i>TTACGTAA</i>	<i>2.64</i>	<i>2.88</i>	[3]
TTACGTCT	2.41	2.07	
TTACGTCG	3.85	4.47	
TGACGCGA	3.94	2.19	CREB_02
TGACGCAC	2.32	3.03	CREB_02
TGACGTAC	1.52	7.06	CREB_02
TGACGTGT	1.53	3.32	CREB_02
TGACGCAT	1.74	3.71	CREB_02
TGACGCAG	1.64	2.00	CREB_02
TTACGTGA	1.76	2.03	
TTACGTAC	1.62	2.54	
TTACGCCA	1.83	2.58	

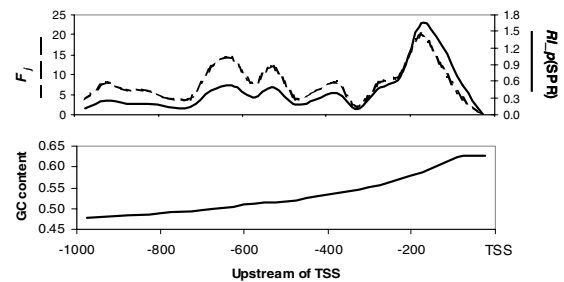


Figure 2. An example of signal enhancement for the motif (TTACGTAA) after considering the GC content of each SPR in the calculation of $RI_p(\text{SPR})$.

4. Future direction

The MotifFilter algorithm is currently being applied to a broader survey of human *cis*-regulatory motifs. Also a comparison with other existing motif prediction methods is underway.

5. Acknowledgement

This is publication NRC 47133 of National Research Council of Canada.

6. References

- [1] Krogh, A. et al. (1994) *J. Mol. Biol.* 235: 1501-1531.
- [2] Matys, V. et al. (2003) *Nucleic Acids Res.* 31:374-378.
- [3] Benbrook, D. M. and N. C. Jones (1994) *Nucleic Acids Res.* 22: 1463-1469.