

Additional materials and results for MUSA

Nuno D. Mendes^a, Ana C. Casimiro^a, Pedro M. Santos^b, Isabel Sá-Correia^b, Arlindo L. Oliveira^a, Ana T. Freitas^{a*}

^a INESC-ID, Instituto Superior Técnico, Rua Alves Redol, 9 1000-029 Lisboa, Portugal

^b Biological Sciences Research Group, Centro de Engenharia Biológica e Química, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001, Lisboa, Portugal

1 APPENDIX A

In this appendix, we present additional material concerned with the methods used and the results obtained.

In particular, we describe the way in which statistical validation of the selected motifs was performed, and the assumptions that were used for this computation.

We also present and discuss the result of applying the proposed method to the artificially generated (synthetic) data sets. The advantage of using synthetic data sets is the ability to specify exactly which motifs we wish to plant in the data against a random background. In this way, we can safely test our method, since we control every aspect of the signal hidden in the random sequences. However, synthetic sequences are still very different from real sequences, since the generators used are not precise enough to model the actual distribution of nucleotides in promoter regions.

This appendix also includes additional results of the application of our method to the data set, composed of 69 putative σ^{54} -dependent promoter sequences of *Pseudomonas putida* KT2440 and the well characterized σ^{54} -dependent promoter *Pu*. These results are very useful to understand the post-processing assembling of motifs presented in table 1 in the original paper.

To conclude we present a comparison of the obtained results with competing approaches. We think that this comparison is very important to understand the main advantages of this new method.

1.1 Statistical significance

In order to assess the statistical significance of motifs we used the method proposed in [3]. This method computes the probability of occurrence, by chance, of a motif composed of two boxes separated by a variable distance in a sequence of a given length. Since the probability of occurrence strongly depends on the overlapping structure of the motif, this method, although not exact, uses generalized geometric approximations in order to calculate the required probability with a good precision. Although the method was proposed to obtain the statistical significance of complex motifs it is easily adapted to assess the significance also of single motifs.

The probability of occurrence of a motif in each sequence is then used to compute the probability of the motif occurring in at least k sequences, by computing the distribution of a sum of Bernoulli variables, each one of them taking the value 1 in accordance with the computed probability for that sequence. This leads to a binomial distribution, if all sequences have equal length, or to a sum of

unequal parameter Bernoulli random variables, if the lengths of the sequences are different.

The statistical significance of each motif is reported by MUSA, and used to order the output.

1.2 Application to synthetic data

The synthetic data sets were produced using a simple random generator based on a first order Markov model to generate nucleotide values outside the planted motifs. Each data set consists of 100 sequences of length 600. The length was chosen to be 600 to conform to the average length of the sequences that will be used in the analysis of real data. These experiments were used mainly to help in the definition of the parameters of the method.

There are two important parameters in our method: λ which defines the length of our λ -mers and ε which defines the tolerance with which we score configurations of λ -mers. Recall that the ε -tolerant score of a configuration, (m_r, m_s, d) , considers the contribution of all configurations (m_r, m_s, d') such that $d' \in [d - \varepsilon, d + \varepsilon]$.

We did not consider the cases where $\lambda \leq 2$ because this will increase the number of most common configurations in each element of the matrix of co-occurrences making the task of identifying motifs harder. The cases in which $\lambda > 4$ have two major drawbacks. Not only the generated matrix is exceedingly large but we will also be unable to identify complex motifs with components shorter than 5 nucleotides or simple motifs less than 6 nucleotides long. Our results only show the cases where $\lambda = 3$ and $\lambda = 4$. We also only considered the cases where $\varepsilon = 0$ and $\varepsilon = 1$ since larger values for our tolerance will inflate the score of most configurations. In any case, we do not discard the interest of performing a broader study.

Experiments with sequences without planted motifs have shown that, for data sets of this approximate size, the algorithms should be able to find and rank motifs of length λ or longer if they occur in at least 20% of the sequences. These conditions are met by the known motifs in the biological data set used.

1.3 Application to the σ^{54} -dependent promoter regions of *P. putida* KT22440

In this section we present results of the application of our method to the data set, composed of 69 putative σ^{54} -dependent promoter sequences of *Pseudomonas putida* KT2440 and the well characterized σ^{54} -dependent promoter *Pu*. In this supplementary data, the motifs that are presented in the following table correspond to the ones that were grouped into families, presented in table 1 in the original paper.

The results presented here were obtained using MUSA and considering the following values for the input parameters: $\lambda = 4$ and

*to whom correspondence should be addressed

$\varepsilon = 1$ (default values). The matrix of motif co-occurrences, in Fig. 1, only considered signals that are present in at least 10% of the input sequences.

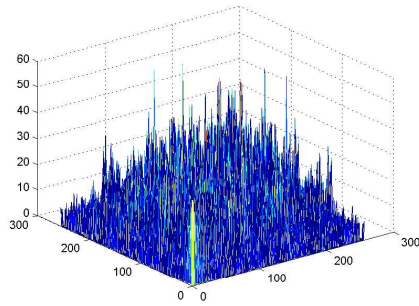


Fig. 1. Matrix of co-occurrences for motifs with quorums above 10% for the σ^{54} -dependent data set

Each of the peaks shown in Fig. 1 corresponds to one motif. Using the bicustering approach these motifs are combined to form larger patterns. The selected motifs were ranked in accordance with their statistical significance. The motifs that have a p-value smaller than 10^{-3} are shown in the following table.

1.4 Comparison with other methods

In this section, we report the results obtained from the application of other methods to the data set composed by the σ^{54} -dependent promoter sequences. We selected three representative methods: MEME [4], a well know algorithm based on EM, commonly used by the scientific community to find common motifs in regulatory regions; MotifSampler [1] a popular Gibbs-sampling based algorithm; and GLAM [2], an alignment based method that works by finding aligned regions in the promoter regions. All methods were used with default values, unless otherwise specified. MEME was instructed to find 10 motifs, with sizes between 3 and 50 nucleotides, and a minimum quorum of 5%. It reported three motifs with significant E-values, shown in table 2. The first consensus reported by MEME is the σ^{54} binding site (families 1, 3, 5 and 8 in table 1 in the original paper). The second and third consensi correspond to family 2 in table in the original paper. The way these consensi are presented is somewhat misleading since not all motifs occur exactly as reported in all sequences and the quorum computed by MEME results from the contributions of all the variations of the reported motifs. For example, if a pattern matching is performed to look for the motif TGGC(7)TTGC, the answer is 36 out of 70 sequences as reported by MUSA. The quorum 69 out of 70 reported by MEME for the first motif of table 2 is obtained allowing variations in different positions of the sequence. MotifSampler was executed with a parameter specifying that it should look for a box with size 8 and all other parameters with default values. It reported in the first place the yTGGCAnn, which corresponds to one box of the σ^{54} binding site. It also reported other consensi, shown in table 3, with significant scores. GLAM, the alignment based method, fails to identify motifs that are not represented in the large majority of sequences.

Number	Motifs	Quorum	p-value
1	TGGC(7)TTGC	36/70	8.8e-37
2	TGGCA(6)TTGC	24/70	7.0e-33
3	TGGC(7)TTGCT	21/70	3.8e-29
4	TTCGCGG(6)CCGC	11/70	3.1e-27
5	TCGCGG(6)CCGC	11/70	3.4e-21
6	TGGCA(6)TTGCT	11/70	8.6e-19
7	CTGGC(7)TTGC	15/70	2.4e-17
8	ATGGCA(6)TTGC	8/70	5.2e-13
9	TGGC(7)TTGCTA	7/70	7.9e-13
10	TGGC(8)TGCTG	9/70	5.3e-09
11	TGGCAC	33/70	3.3e-08
12	CTGGCA	33/70	1.8e-07
13	TGGCA	54/70	2.9e-07
14	CTGGC	54/70	8.5e-07
15	CGCGAAG	14/70	1.3e-06
16	TGGCATGG	9/70	1.0e-05
17	GGCAC	49/70	1.0e-05
18	CCGCTCC	13/70	1.7e-05
19	CTGGCAC	15/70	3.0e-05
20	TAACAAG	9/70	3.7e-05
21	GCTGGC	32/70	4.1e-05
22	AAGGTTT	10/70	0.1e-03
23	TTGGCAC	14/70	0.1e-03
24	GCTGGCA	16/70	0.1e-03
25	TGGCATG	14/70	0.1e-03
26	AAACCC	22/70	0.1e-03
27	CAAAACCC	7/70	0.2e-03
28	CGCGAA	21/70	0.3e-03
29	ACAAGAA	9/70	0.3e-03
30	GGTTTT	20/70	0.3e-03
31	TCAGTG	17/70	0.4e-03
32	TTTTAT	14/70	0.4e-03
33	TTGGCA	26/70	0.6e-03
34	GGCACAGC	7/70	0.7e-03
35	GAGCGGG	10/70	0.8e-03
36	TGGCAT	22/70	0.9e-03
37	GCCTGT	21/70	0.9e-03

Table 1. Motifs reported by MUSA with p-value smaller than $10e - 3$ for the σ^{54} -dependent data set

As such, it reports only one motif with a significant p-value, the σ^{54} consensus.

The comparative analysis carried out with other methods shows that MUSA is very competitive in finding and reporting, in a readable way, motifs that are of biological significance. While all the other commonly used methods are also able to find motifs that correspond to very strong signals, MUSA is unique in its ability to find lower quorum motifs and to report them in a clear and easy to understand way. Additionally, MUSA requires the specification of only a minimal number of parameters, and, in most cases, can be used using all the default values.

2 APPENDIX B

In this appendix we present the algorithmic details of the procedures used to generate the matrix of co-occurrences and to extract

Number	Multilevel Consensi	Quorum	E-value
1	GCTGGCACGGCTCTTGCT T TACGCG TG	69/70	5.5e-113
2	CCTCTTCGCGGGTAAACCGCTCCTACAG GG G CGCG C GA C	15/70	6.4e-63
3	GGCCCCCTCGCGGGCAAGCCCGCTCCCAC C TTT A T TA G T	8/70	2.4e-16

Table 2. Multilevel consensi reported by the MEME program for the σ^{54} -dependent data set

Consensi	Score
yTGGCAnn	404.4
CyCCnnmA	310.8
swnwnCCn	345.6
yTnTGnnk	299.3
rGsCmTTG	317.0
CrnGnAAA	310.0

Table 3. Consensi reported by the MotifSampler program for the σ^{54} -dependent data set

the biclusters. Algorithm 1 computes the ε -tolerant matrix of co-occurrences for a given set of sequences \mathcal{S} . Its inputs are the set of sequences \mathcal{S} , the value λ which defines the length of the λ -mers that will be considered, and ε which defines the tolerance. $\text{Occ}_i(m)$ refers to the list of occurrences of the λ -mer m in the i th sequence of \mathcal{S} . Conf_i is a set of configurations of pairs of λ -mers being computed for the i th sequence. $\text{Score}[(m_r, m_s, d)]$ can be seen as a property of a particular configuration of λ -mers. And, finally, $M[r, s]$ is the element of the matrix of co-occurrences which keeps the score of the most common configuration of the r th and s th λ -mers in $L(\mathcal{S})$. Algorithm 2 extracts biclusters from a matrix of co-occurrences. Its inputs are a matrix of co-occurrences \mathcal{M} and a value minscore that defines the minimum score required for matrix elements to be used to start generating a bicluster. The algorithm considers decreasing values of possible scores h , starting with the maximum possible value $|\mathcal{S}|$ down to the specified minimum value minscore. For each score value h , the algorithm keeps a set of biclusters biclusters_h . At the end we obtain $(|\mathcal{S}| - \text{minscore} + 1)$ such sets.

REFERENCES

- [1] Thijs G., Marchal K., Lescot M., Rombauts S., De Moor B., Rouze P., and Moreau Y. A gibbs sampling method to detect over-represented motifs in upstream regions of coexpressed genes. *Journal of Computational Biology*, 9(2):447–464, 2002.
- [2] Frith M., Hansen U., Spouge J., and Weng Z. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Research*, 32(1):189–200, 2004.
- [3] Robin S., Daudin J-J, Richard H., Sagot M-F, and Schbath S. Occurrence probability of structured motifs in random sequences. *Journal of Computational Biology*, 9(6):761–774, 2002.
- [4] Bailey T. L. and Elkan C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21(1-2):51–80, 1995.

Algorithm 1 Computes the ε -tolerant matrix of co-occurrences

```

1: for all  $m_r, m_s \in L(\mathcal{S})$  do
2:   MaxScore  $\leftarrow 0$ 
3:   for all  $S_i \in \mathcal{S}$  do
4:     Confi  $\leftarrow \emptyset$ 
5:     for all  $c_r \in \text{Occ}_i(m_r), c_s \in \text{Occ}_i(m_s)$  do
6:        $d \leftarrow c_r - c_s$ 
7:       if  $d \neq 0$  then
8:         Confi  $\leftarrow \text{Conf}_i \cup \{(m_r, m_s, d)\}$ 
9:         for  $k = 1$  to  $\varepsilon$  do
10:           if  $d + k \neq 0$  then
11:             Confi  $\leftarrow \text{Conf}_i \cup (m_r, m_s, d + k)$ 
12:           end if
13:           if  $d - k \neq 0$  then
14:             Confi  $\leftarrow \text{Conf}_i \cup (m_r, m_s, d - k)$ 
15:           end if
16:         end for
17:       end if
18:     end for
19:   for all  $(m_r, m_s, d) \in \text{Conf}_i$  do
20:     Score[( $m_r, m_s, d$ )]  $\leftarrow \text{Score}[(m_r, m_s, d)] + 1$ 
21:   end for
22: end for
23: for all  $(m_r, m_s, d) \in \bigcup_i \text{Conf}_i$  do
24:   if Score[( $m_r, m_s, d$ )] > MaxScore then
25:     MaxScore  $\leftarrow \text{Score}[(m_r, m_s, d)]$ 
26:   end if
27: end for
28:  $M[r, s] \leftarrow \text{MaxScore}$ 
29: end for

```

Algorithm 2 Extracts biclusters in a matrix of co-occurrences

```

1: for  $h = |\mathcal{S}|$  to minscore do
2:   biclustersh  $\leftarrow \emptyset$ 
3:   for all  $a_{ij} \in \mathcal{M}$  with score  $h$  do
4:     if  $a_{ij} \notin \bigcup_{B_k \in \text{biclusters}_h} B_k$  then
5:       if  $i = j$  then
6:          $I \leftarrow \{i\}$ 
7:          $\Lambda \leftarrow \{i\}$ 
8:       else
9:          $I \leftarrow \{i, j\}$ 
10:         $\Lambda \leftarrow \emptyset$ 
11:      end if
12:      for  $k = 1$  to  $|L(\mathcal{S})|$  do
13:        if  $B(I \cup \{k\}, \Lambda)$  is  $h$ -valid then
14:           $I \leftarrow I \cup \{k\}$ 
15:          if  $B(I, \Lambda \cup \{k\})$  is  $h$ -valid then
16:             $\Lambda \leftarrow \Lambda \cup \{k\}$ 
17:          end if
18:        end if
19:      end for
20:      biclustersh  $\leftarrow \text{biclusters}_h \cup \{B(I, \Lambda)\}$ 
21:    end if
22:  end for
23: end for

```