Finding functional promoter motifs by computational methods: a word of caution

Rajesh Chowdhary

Institute for Infocomm Research,
21 Heng Mui Keng Terrace, Singapore 119613

School of Computing,
National University of Singapore, Singapore 117543
E-mail: rajesh@i2r.a-star.edu.sg

Limsoon Wong

School of Computing,
National University of Singapore, Singapore 117543
E-mail: wongls@comp.nus.edu.sg

Vladimir B. Bajic*

Institute for Infocomm Research,
21 Heng Mui Keng Terrace, Singapore 119613

South African National Bioinformatics Institute,
University of the Western Cape,
Private Bag X17, Bellville 7535, South Africa
E-mail: vlad@sanbi.ac.za
*Corresponding author

Abstract: The standard practice in the analysis of promoters is to select promoter regions of convenient length. This may lead to false results when searching for Transcription Factor Binding Sites (TFBSs), since the sequences may contain coding segments. In such cases, motif detection may single out motifs from the coding regions. The mapping of TFBSs to promoters may result in a misleading picture of 'promoter' content. We illustrate these issues using the example of histones H2A and H2B and show how such analysis could be misleading if care is not exercised to eliminate coding regions from the presumed promoter sequences.

Keywords: promoter; Transcription Start Sites (TSSs); Translation Initiation Sites (TIS); Transcription Factor Binding Sites (TFBSs); motifs.

Reference to this paper should be made as follows: Chowdhary, R., Wong, L. and Bajic, V.B. (2006) 'Finding functional promoter motifs by computational methods: a word of caution', *Int. J. Bioinformatics Research and Applications*, Vol. 2, No. 3, pp.282–288.

Biographical notes: Rajesh Chowdhary received his BTech from the Indian Institute of Technology, Bombay, and MSc from Imperial College, London. Currently, he is pursuing his PhD in Computer Science from the National University of Singapore. His research interests include Bayesian networks, biomedical informatics, machine learning, gene regulation, and genome data mining.

Limsoon Wong is a Professor of Computing and of Medicine at the National University of Singapore. He is currently working mostly on knowledge discovery technologies and is especially interested in their application to biomedicine. Limsoon has written about 100 research papers, a few of which are among the best cited of their respective fields. He serves on the editorial boards of several journals, and is a scientific advisor to a number of companies.

Vladimir B. Bajic is Professor in the South African National Bioinformatics Institute (SANBI), University of the Western Cape (South Africa). He is a member of the Academy of Nonlinear Sciences (Russia). His current interest is in biological discoveries through applications of artificial intelligence for complex biomedical problems. He has contributed more than 300 research papers, software, and edited volumes, and has designed numerous bioinformatics systems. He serves on the editorial boards of several journals.

1 Introduction

Eukaryotic promoter content analysis frequently involves detection of putative TFBSs in a set of co-regulated genes (Claverie and Sauvaget, 1985; Wasserman and Fickett, 1998; Krivan and Wasserman, 2001). A common approach involves the selection of an arbitrary segment of flanking sequence around the Transcription Start Site (TSS) of the target genes and the application of computational methods that search for putative TFBSs in the assumed promoter region. Examples of selecting presumed promoter segments for analysis include studies by FitzGerald et al. (2004) on region [–2500, +500], by Marino-Ramirez et al. (2004) on region [+2000, –1000], by Zhang et al. (2002) on region [–2000, –1], by Zheng et al. (2003) on region [–600, –1], by Bajic et al. (2004) on region [–70, +60], by Kel-Margoulis et al. (2003) on region [–300, +50], by Prakash et al. (2004) on region [–1000, –1] and by Frith et al. (2001, 2004) where regions considered were [–249, +50], [–1499, +500] and [–1500, –1] with respect to the TSS; Blanchette et al. (2003) analysed promoters in the range 250–2000 bp upstream of the TSS, and Podvinec et al. (2002) analysed assumed promoter regions of various sizes in the range 3400 bp and 28600 bp upstream of TSS. The large variations in the assumed promoter regions used for promoter motif studies may be because of the fact that the promoter regions (core, proximal and distal) may vary a lot from one gene to the other. Added to this, functionally important TFBSs are non-uniformly distributed along the DNA and are not specific to a particular region. For example, TFBS may be found upstream of the TSS, in the 5′ UTR (Butler and Kadonaga, 2002), in the first intron (Wasserman and Sandelin, 2004) and further downstream to it. Thus, promoter segments chosen for analysis generally depend on one's experimental objectives and convenience.

The methodology of arbitrary selection of promoter region boundaries may, however, be risky and can lead to misleading results about what potential TFBSs and other potentially important promoter motifs are. As can be seen from the references above, the practice of utilising presumed promoter regions of different length for various analyses is quite common. What is missing, however, is that these references do not make any reference to neither the elimination of the upstream coding segments that are part of the neighbouring gene nor the elimination of the downstream coding segments when promoter region extends downstream of TSS. By selecting boundaries of promoters for motif analysis, we implicitly assume that the whole selected region actually belongs to the gene's promoter and thus harbours TFBSs. However, this assumption may not be correct if the genomic equivalent of Translation Initiation Site (geTIS) is close to TSS (the case of short 5′ UTRs) and is part of arbitrarily selected segments around the TSS, which makes genes' coding regions part of these segments. Such a situation can arise when the analysed segment overlaps either with the coding region of the same gene downstream of the TSS, or with the coding region of a different gene located nearby upstream of the TSS on the opposite strand (divergently transcribed gene).

The accidental presence of the coding regions in the candidate promoter segments may lead to incorrect promoter motif analysis. If the coding regions are more conserved than the real promoter regions in the candidate segments, motif programs may be biased towards detecting motifs from the coding regions that are likely to have higher significance compared to motifs from the actual promoter region. Since programs that detect motifs usually rely on some statistical significance criteria, motifs with higher significance are preferred, while those with lower significance are filtered out. Also, the motifs discovered from the coding regions may wrongly be considered legitimate putative TFBSs, making the analysis incorrect. The presence of coding regions in the analysed promoter segments, thus, may affect the quality and accuracy of the results.

In this note, we illustrate the above-mentioned issues by considering an example of divergently transcribed human histone H2A and H2B genes that share a common promoter. We show that motif discovery results could be deceptive if the analysed promoter sequences contain the coding regions. We highlight the importance of being aware of the boundaries of the actual promoter regions so that coding regions are not included in the promoter sequences used for motif analysis.

2 Experiment summary and results

While performing a motif analysis at the 5′ end region [–1000, +500] of 14 human histone H2A-H2B genes, we observed that that the motifs discovered by a motif discovery program MEME (Bailey and Elkan, 1994)/MAST (Bailey and Gribskov, 1998) were almost confined to those regions of the candidate segments that represented the coding regions (see Figure 1). This observation that the promoter in these genes was devoid of TFBSs apparently seemed erroneous and led us to investigate further. Subsequently, we separately investigated segments that represented greater parts of the actual promoter regions [–250, –1], but contained no coding regions. We maintained the same MEME/MAST parameters for both the experiments. We denote the two analysed regions as Long Segment (LS) [–1000, +500] and Real Promoter (RP) [–250, –1].

The gene set that we analysed contained seven H2A-H2B gene pairs, namely, H2A/a-H2B/a, H2A/c-H2B/c, H2A/d-H2B/d, H2A/e-H2B/e, H2A/g-H2B/g, H2A/l-H2B/l and H2A/n-H2B/n. Each pair represents divergently transcribed genes on opposite strands of DNA that share a bidirectional promoter that is generally less than 320 bp in length (Albig et al., 1999; Trappe et al., 1999) and contain many TFBSs.

Out of 100 motifs returned by MEME, MAST selected 48 motifs for LS sequence analysis and six motifs for RP sequence analysis based on their statistical significance using cut-off *E* value less than 1. Figures 1 and 2 show motif distributions returned by MAST for LS and RP sequences, respectively.

Figure 1 Motif distribution obtained from MAST in LS genomic sequences [–1000, +500] of 14 human histone H2A and H2B genes. We observe two motif clusters, Cluster I and Cluster II, and the promoter region between them devoid of any motif. Motifs shown above were detected in LS by MEME. (+/–) sign with motif indicates the strand. Gene is presented in the format ‘species|histone_group|geneID|strand|chromosome|official_name|alternative_name’. We used MEME and MAST programs to discover motifs in sequences of LS and RP data sets. Motifs discovered by MEME in the data sets were used by MAST to evaluate the presence of combined motif patterns in the same sets of sequences. Parameters used for MEME were zoops model, motif width from 6 to 12, maximum number of motifs to search as 100 and reverse complimentary strand was considered; parameters used for MAST were all significant motifs (with motif *E* value less than 1.0) returned by MEME, motif was reported if its sequence *p* value was less than 0.005, correlated motifs were filtered out, both strands were searched and individual sequence composition was used to calculate *p* and *E* values

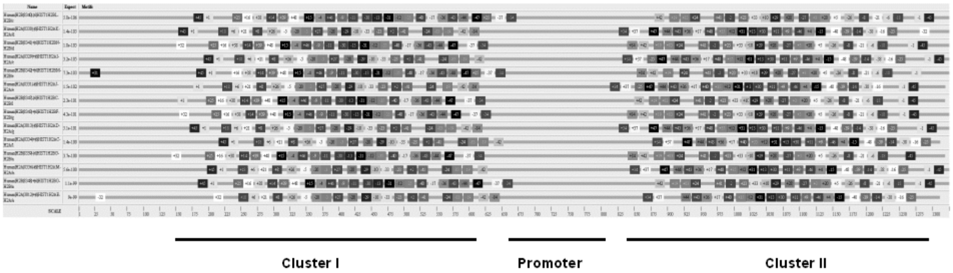
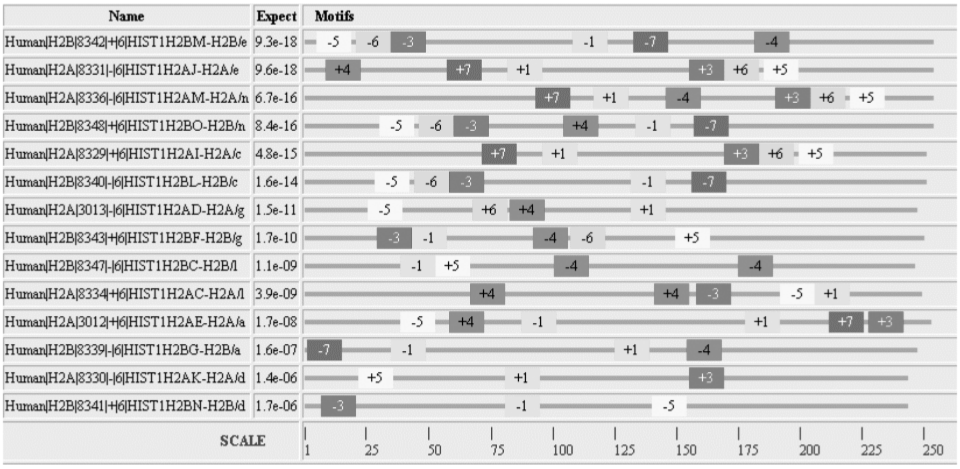


Figure 2 Motif distribution obtained from MAST in RP genomic sequences [–250, –1] of 14 human histone H2A and H2B genes. Motifs shown above were detected in RP by MEME. (+/–) sign with motif indicates the strand. Gene is presented in the format ‘species|histone_group|geneID|strand|chromosome|official_name|alternative_name’



<table>
<tr><th>Name</th><th>Expect</th><th>Motifs</th></tr>
<tr><td>Human|H2B|8342|+|6|HIST1H2BM-H2B/e</td><td>9.3e-18</td><td>-5 -6 -3 -1 -7 -4</td></tr>
<tr><td>Human|H2A|8331|-|6|HIST1H2AJ-H2A/e</td><td>9.6e-18</td><td>+4 +7 +1 +3 +6 +5</td></tr>
<tr><td>Human|H2A|8336|-|6|HIST1H2AM-H2A/n</td><td>6.7e-16</td><td>+7 +1 -4 +3 +6 +5</td></tr>
<tr><td>Human|H2B|8348|+|6|HIST1H2BO-H2B/n</td><td>8.4e-16</td><td>-5 -6 -3 +4 -1 -7</td></tr>
<tr><td>Human|H2A|8329|+|6|HIST1H2AI-H2A/c</td><td>4.8e-15</td><td>+7 +1 +3 +6 +5</td></tr>
<tr><td>Human|H2B|8340|-|6|HIST1H2BL-H2B/c</td><td>1.6e-14</td><td>-5 -6 -3 -1 -7</td></tr>
<tr><td>Human|H2A|3013|-|6|HIST1H2AD-H2A/g</td><td>1.5e-11</td><td>-5 +6 +4 +1</td></tr>
<tr><td>Human|H2B|8343|+|6|HIST1H2BF-H2B/g</td><td>1.7e-10</td><td>-3 -1 -4 -6 +5</td></tr>
<tr><td>Human|H2B|8347|-|6|HIST1H2BC-H2B/l</td><td>1.1e-09</td><td>-1 +5 -4 -4</td></tr>
<tr><td>Human|H2A|8334|+|6|HIST1H2AC-H2A/l</td><td>3.9e-09</td><td>+4 +4 -3 -5 +1</td></tr>
<tr><td>Human|H2A|3012|+|6|HIST1H2AE-H2A/a</td><td>1.7e-08</td><td>-5 +4 -1 +1 +7 +3</td></tr>
<tr><td>Human|H2B|8339|-|6|HIST1H2BG-H2B/a</td><td>1.6e-07</td><td>-7 -1 +1 -4</td></tr>
<tr><td>Human|H2A|8330|-|6|HIST1H2AK-H2A/d</td><td>1.4e-06</td><td>+5 +1 +3</td></tr>
<tr><td>Human|H2B|8341|+|6|HIST1H2BN-H2B/d</td><td>1.7e-06</td><td>-3 -1 -5</td></tr>
<tr><td>SCALE</td><td></td><td>1 25 50 75 100 125 150 175 200 225 250</td></tr>
</table>

We observe that 48 LS sequence motifs (Figure 1) form two large clusters on both sides of TSS. Further investigation reveals that these two clusters correspond to the coding regions of H2A-H2B genes. While in H2A (H2B) the cluster on the downstream side represents its own coding region, the cluster on the upstream side represents the coding region of its divergently transcribed gene pair H2B (H2A). This is also evident from motif distribution of H2A which is the mirror image of its corresponding H2B pair with strands opposite, and vice versa. Notably, the actual promoter regions of these genes that lie between the two clusters were completely devoid of motifs (Figure 1). This might be because H2A and H2B are evolutionarily conserved proteins (Luo and Dean, 1999; Doenecke et al., 1997), and therefore the genomic regions corresponding to these proteins should contain conserved motifs, likely better conserved than motifs from the actual promoter in statistical terms. In such cases, analysis could be biased in favour of the motifs from the coding regions compared to those from the promoter regions. Such situations may appear because motif discovery programs make no distinction between coding and promoter regions and generally detect motifs that qualify predefined statistical significance threshold levels.

With the same MEME/MAST parameter settings as LS, our analysis on RP sequences which contained greater parts of H2A-H2B promoters and no coding regions, however, supports known experimental results to a great extent. We observed that most of the motifs detected in the RP sequences (Figure 2) correspond well with the experimentally verified motifs in H2A and H2B genes (Albig et al., 1999; Trappe et al., 1999): Motif 1 (GCCCAATCAAAA) corresponds to CCAAT-box, Motif 4 (ATGCAAATGAGG) represents Oct-1 box and Motif 5 (GCTATAAAATGC) represents TATA-box, while Motif 3 (AGCTTCCTTTTC) and Motif 7 (GATGACGACAG) partially represent E2F-box and CRE-box, respectively.

Our key observation from the above analysis is that promoter region motifs that do not appear detected in the LS analysis appear clearly in the analysis of RP. This suggests that the selection of the region around the TSS for the analysis of promoter content is critical and must be carefully done, so as not to contain parts of the coding region of the analysed and neighbouring genes. We have noted in our analysis that neighbouring genes may be located critically close to each other. If care is not exercised in the region selection, the motif discovery could be strongly affected, as shown in this analysis. The presence of the coding regions in the analysed promoter segment may thus potentially result in

- putative TFBS motifs of the promoter region that may not be detected, as shown in our example
- motifs of the coding region that being detected and apparently, by error, could be considered as putative TFBS motifs.

Therefore, while selecting promoter segments for motif analysis due care should be given to

- the distance between gene’s TSS and its geTIS (i.e., 5′ UTR)
- the distance between geTISs of any potential divergently transcribed genes (neighbouring genes).

By doing so we may be able to avoid coding regions that might potentially overlap with the analysed promoter regions. Coding regions downstream of the geTISs may be less of a problem as geTISs are generally well annotated. However, interference of the coding regions of neighbouring genes upstream of the TSS may not be so intuitive and hence can be troublesome.

References

Albig, W., Trappe, R., Kardalinou, E., Eick, S. and Doenecke, D. (1999) 'The human H2A and H2B histone gene complement', *Biol. Chem.*, Vol. 380, Vol. 1, pp.7–18.

Bailey, T.L. and Elkan, C. (1994) 'Fitting a mixture model by expectation maximization to discover motifs in biopolymers', *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, Vol. 2, pp.28–36.

Bailey, T.L. and Gribskov, M. (1998) 'Combining evidence using p-values: application to sequence homology searches', *Bioinformatics*, Vol. 4, pp.48–54.

Bajic, V.B., Choudhary, V. and Hock, C.K. (2004) 'Content analysis of the core promoter region of human genes', *In Silico Biol.*, Vol. 4, No. 2, pp.109–125.

Blanchette, M., Kwong, S. and Tompa, M. (2003) 'An empirical comparison of tools for phylogenetic footprinting', *Third IEEE Symposium on Bioinformatics and Bioengineering*, pp.69–78.

Butler, J.E. and Kadonaga, J.T. (2002) 'The RNA polymerase II core promoter: a key component in the regulation of gene expression', *Genes Dev.*, Vol. 16, No. 20, pp.2583–2592.

Claverie, J.M. and Sauvaget, I. (1985) 'Assessing the biological significance of primary structure consensus patterns using sequence databanks. I. Heat-shock and glucocorticoid control elements in eukaryotic promoters', *Comput. Appl. Biosci.*, Vol. 1, No. 2, pp.95–104.

Doenecke, D., Albig, W., Bode, C., Drabent, B., Franke, K., Gavenis, K. and Witt, O. (1997) 'Histones: genetic diversity and tissue-specific gene expression, a review', *Histochem. Cell Biol.*, Vol. 107, No. 1, pp.1–10.

FitzGerald, P.C., Shlyakhtenko, A., Mir, A.A. and Vinson, C. (2004) 'Clustering of DNA sequences in human promoters', *Genome Res.*, Vol. 14, No. 8, pp.1562–1574, Epub.

Frith, M.C., Hansen, U. and Weng, Z. (2001) 'Detection of cis-element clusters in higher eukaryotic DNA', *Bioinformatics*, Vol. 17, No. 10, pp.878–889.

Frith, M.C., Fu, Y., Yu, L., Chen, J.F., Hansen, U. and Weng, Z. (2004) 'Detection of functional DNA motifs via statistical over-representation', *Nucleic Acids Res.*, Vol. 32, No. 4, pp.1372–1381.

Kel-Margoulis, O.V., Tchekmenev, D., Kel, A.E., Goessling, E., Hornischer, K., Lewicki-Potapov, B. and Wingender, E. (2003) 'Composition-sensitive analysis of the human genome for regulatory signals', *In Silico Biol.*, Vol. 3, Nos. 1–2, pp.145–171.

Krivan, W. and Wasserman, W.W. (2001) 'A predictive model for regulatory sequences directing liver-specific transcription', *Genome Res.*, Vol. 11, No. 9, pp.1559–1566.

Luo, R.X. and Dean, D.C. (1999) 'Chromatin remodeling and transcriptional regulation', *J. Natl. Cancer. Inst.*, Vol. 91, No. 15, pp.1288–1294.

Marino-Ramirez, L., Spouge, J.L., Kanga, G.C. and Landsman, D. (2004) 'Statistical analysis of over-represented words in human promoter sequences', *Nucleic Acids Res.*, Vol. 32, No. 3, pp.949–958.

Podvinec, M., Kaufmann, M.R., Handschin, C. and Meyer, U.A. (2002) 'NUBIScan, an in Silico approach for prediction of nuclear receptor response elements', *Mol. Endocrinol.*, Vol. 16, No. 6, pp.1269–1279.

Prakash, A., Blanchette, M., Sinha, S. and Tompa, M. (2004) 'Motif discovery in heterogeneous sequence data', *Pac. Symp. Biocomput.*, pp.348–359.

Trappe, R., Doenecke, D. and Albig, W. (1999) 'The expression of human H2A-H2B histone gene pairs is regulated by multiple sequence elements in their joint promoters', *Biochim. Biophys. Acta.*, Vol. 1446, No. 3, pp.341–351.

Wasserman, W.W. and Fickett, J.W. (1998) 'Identification of regulatory regions which confer muscle-specific gene expression', *J. Mol. Biol.*, Vol. 278, No. 1, pp.167–181.

Wasserman, W.W. and Sandelin, A. (2004) 'Applied bioinformatics for the identification of regulatory elements', *Nat. Rev. Genet.*, Vol. 5, No. 4, pp.276–287.

Zhang, H., Ramanathan, Y., Soteropoulos, P., Recce, M.L. and Tolias, P.P. (2002) 'EZ-retrieve: a web-server for batch retrieval of coordinate-specified human DNA sequences and underscoring putative transcription factor-binding sites', *Nucleic Acids Res.*, Vol. 30, No. 21, p.e121.

Zheng, J., Wu, J. and Sun, Z. (2003) 'An approach to identify over-represented cis-elements in related sequences', *Nucleic Acids Res.*, Vol. 31, No. 7, pp.1995–2005.