A mathematical method for the classification of promoter sequences from the *A.thaliana* genome

M.A. Korotkova, A.M.Kamionskya and E.V. Korotkov

1National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, Russian Federation
2Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences, Moscow, Russia.

genekorotkov@gmail.com

Abstract. A mathematical method for creating classes of promoter sequences has been developed. The method was used to calculate the classes of promoter sequences from the *A.thaliana* genome. A total of 16 statistically significant classes of promoter sequences were obtained with class sizes ranging from 8,000 to 100 promoters. The classes obtained allow us to identify potential promoter sequences in various genomes with the number of false positives not exceeding $10^3$ per genome.

1. Introduction

The promoter sequence in both prokaryotes and eukaryotes is located around the transcription start point [1]. This point is the first base of the first codon (denoted as +1) of the gene coding sequence. For eukaryotic genomes, the promoter region captures the sequence from -499 to +100 and this region has a length of 600 nucleotides, although it is difficult to identify strict boundaries for the promoter sequences. Further, we will focus only on eukaryotic promoter sequences. The promoter includes some motifs, which are short conservative sequences. The so-called TATA sequence is known, which occupy positions from -31 to -26 nucleotides. Also known is the B recognition element (BRE), which is between -37 and 32 nucleotides in the promoter sequence. Short sequences have also been identified that allow the binding of various protein factors to the promoter sequence. Many of these sequences fall on the promoter region from +1 to +40. The promoter sequence is not symmetrical and this will cause the DNA polymerase to begin transcription in the right direction. The frequency of gene transcription depends on which promoter is located near the gene.

By activity, promoters can be divided into two classes. The first class includes constitutive promoters that provide a constant level of transcription. Inducible promoters can be attributed to the second class. In this case, the transcription rate depends on the presence of certain transcription factors or other external conditions. For inducible promoters, the transcription frequency may be zero, in which case the gene is in standby mode. Promoter sequences are very different from each other. This is due to the need to control the transcription rate of various genes. When transcription is initiated, a transcriptional complex is assembled on the promoter sequence. The complex includes RNA polymerase
and dozens of other transcription factors. The set of these factors can vary from one gene to another, which in turn leads to a strong variety of promoter sequences.

Today, hundreds of thousands of promoter sequences from various eukaryotic genomes are known. Databases have been created for promoter sequences. In this work, we used the EPD database, which is located at https://epd.epfl.ch//index.php. However, despite a large number of promoter sequences, it has not yet been possible to find a statistically significant multiple alignment between them. This led to the fact that identifying (annotation) promoters by nucleotide sequence is quite difficult. A typical annotation scheme is to build a statistically significant multiple alignment. Then this alignment is used for profile analysis or for constructing a hidden Markov model. Such a scheme leads to a low number of false positives in the analysis of genomes. However, algorithms for predicting promoter sequences currently use other mathematical methods due to the lack of statistically significant multiple alignment. These are such algorithms as TSSW [2], PePPER [3], G4PromFinder [4] and many others. As a result, on average, different algorithms predict one promoter (false positive) for several thousand bases of DNA, while in the human genome there are $10^5$ bases per gene. As a result, it is impossible to distinguish the true promoter from false predictions. It turns out that to use mathematical methods to search of promoters, we must reduce the number of false positives by about $10^2$ or $10^3$ times.

If we can correctly predict promoter sequences, we can more accurately identify transcript star sites (TSS) and genes.

We previously developed a mathematical method for creating multiple alignment for very different nucleotide sequences (MAHDS). Multiple alignment for nucleotide sequences can be calculated at http://victoria.biengi.ac.ru/mahds/auth. We will understand by highly differing sequences those sequences that have accumulated more than 2.5 random substitutions ($x$) for one nucleotide relative to each other. MAHDS constructs statistically significant alignments for $x$ in the range from 2.5 to 4.4. We have shown that previously developed algorithms can calculate statistical significant multiple alignments for $x < 2.5$ [4]. We used MAHDS to construct multiple alignment of promoter sequences from the A.thaliana genome. Previously, statistically significant multiple alignment was not constructed for promoter sequences, since for them $x = 3.7$. This follows from the alignment of 100 artificial sequences with a length of 600 bases with different [5].

2. Results and discussion

Promoter sequences were obtained from the site: https://epd.epfl.ch//index.php [6]. We used 22694 promoter sequences from the A.Thaliana genome. Denote the set of these sequences as $S$. Each of the promoters had a length equal to 600 nucleotides. The promoter included sequences from -499 to +100 relative to the first base of the first codon (+1 gene position). To create classes of promoter sequences, we first used the MAHDS method to construct multiple alignment of 22694 sequences.

To perform multiple alignment, we united the promoter sequences into one $L$ sequence (Figure 1, point 1). We aligned the sequence $L$ with respect to position weighted matrices [7] using the dynamic programming. Position weight matrices were optimized by the genetic algorithm [8]. The purpose of this optimization was to find a position weight matrix $W(600,16)$ that has the greatest value of the similarity function $F_{max}$ (Figure 1, point 2) [9]. Here 600 is the length of the promoter sequences, 16 is the number of base pairs. We used a position weight matrix, where there were 16 rows. Each row contained the weight of a base pair at positions $i$ and $i+1$, where $i$ changed from 1 to 599.

As a result, we obtained multiple alignment for which we calculated the best positional weight matrix $W(600,16)$. However, only a part of the promoter sequences from the set $S$ can have a statistically significant alignment with the matrix $W(600,16)$. Therefore, we aligned each promoter sequence from the set $S$ with respect to the matrix $W(600,16)$ (Figure 1, point 3) [10]. In the class, we selected such promoter sequences from the set $S$ that have nonrandom alignment with respect to the matrix $W(600,16)$. Such promoter sequences can be assigned to one class. This class is characterized by the best matrix $W(600,16)$ (Figure 1, point 4).
To create further classes, we removed from the set $S$ all sequences that have nonrandom alignment with respect to the matrix $W(600,16)$ and obtained the set $S_1$ (Figure 1, point 6). Further, for $S_1$, the procedure for creating the class was iteratively repeated, as was done above for the set $S$. If during the iterative procedure the volume of the set $S$ became less than 100 promoters, then the classification procedure stopped.

Table 1. No. shows the class number; $V$ shows the volume of the class among the promoters.

| No | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|    | 2419| 1071| 1275| 355 | 387 | 421 | 317 | 156 | 217 |

In total, we managed to create 25 statistically significant classes of promoter sequences with class volumes ranging from 8000 to 100. These 25 classes combined more than 75% of the promoter sequences. Table 1 shows the size of the first 10 classes. The remaining classes contain less than 200 but more than 100 promoters. The results show that it is possible to calculate the statistical important multiple alignment of promoter sequences.
The promoter sequences of each class were used to search for promoters in the A. thaliana genome. It was possible to show that in this case one false positive occurs in approximately $10^7$ nucleotides. This means that classes of promoters can be used to search for promoter sequences in eukaryotic genomes.

This work was supported by the grant of Russian Fund of Fundamental Investigations, N.20-016-00057A

References
[1] Levine M and Tjian R 2003 Transcription regulation and animal diversity *Nature* 424 147–51
[2] Solovyev V V., Shahmuradov I A and Salamov A A 2010 Identification of promoter regions and regulatory sites. *Methods in molecular biology (Clifton, N.J.)* 674 57–83
[3] de Jong A, Pietersma H, Cordes M, Kuipers O P and Kok J 2012 PePPER: a webserver for prediction of prokaryote promoter elements and regulons *BMC Genomics* 13 299
[4] Di Salvo M, Pinatel E, Talà A, Fondi M, Peano C and Alifano P 2018 G4PromFinder: An algorithm for predicting transcription promoters in GC-rich bacterial genomes based on AT-rich elements and G-quadruplex motifs *BMC Bioinformatics* 19 36
[5] Korotkov E V, Suvorova Y M, Kostenko D O and Korotkova M A 2020 Multiple alignment of promoter sequences from the A.thaliana genome *Neural Computing and Applications*
[6] Dreos R, Ambrosini G, Groux R, Cavin Périer R and Bucher P 2017 The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms *Nucleic Acids Research* 45 D51–5
[7] Laskin A A, Korotkov E V, Chaley M B and Kudryashov N A 2003 The Locally Optimal Method of Cyclic Alignment to Reveal Latent Periodicities in Genetic Texts: The NAD-binding Protein Sites *Molecular Biology* 37
[8] Pugacheva V, Korotkov A and Korotkov E 2016 Search of latent periodicity in amino acid sequences by means of genetic algorithm and dynamic programming *Statistical Applications in Genetics and Molecular Biology* 15 381–400
[9] Korotkov E V. and Korotkova M A 2017 Search for regions with periodicity using the random position weight matrices in the C. elegans genome *International Journal of Data Mining and Bioinformatics* 18 331
[10] Laskin A A, Korotkov E V., Chaley M B and Kudryashov N A 2003 The locally optimal method of cyclic alignment to reveal latent periodicities in genetic texts. The NAD-binding protein sites *Molekuliarnaia Biologiiia* 37 663–73