International Journal of
Molecular Sciences

MDPI

Article

Classification of Promoter Sequences from Human Genome

Konstantin Zaytsev [1], Alexey Fedorov [1] and Eugene Korotkov [2,*]

[1] Bach Institute of Biochemistry, Federal Research Center of Biotechnology of the Russian Academy of Sciences, 119071 Moscow, Russia
[2] Institute of Bioengineering, Federal Research Center of Biotechnology of the Russian Academy of Sciences, 119071 Moscow, Russia
* Correspondence: katrin2@biengi.ac.ru

Abstract: We have developed a new method for promoter sequence classification based on a genetic algorithm and the MAHDS sequence alignment method. We have created four classes of human promoters, combining 17,310 sequences out of the 29,598 present in the EPD database. We searched the human genome for potential promoter sequences (PPSs) using dynamic programming and position weight matrices representing each of the promoter sequence classes. A total of 3,065,317 potential promoter sequences were found. Only 1,241,206 of them were located in unannotated parts of the human genome. Every other PPS found intersected with either true promoters, transposable elements, or interspersed repeats. We found a strong intersection between PPSs and Alu elements as well as transcript start sites. The number of false positive PPSs is estimated to be 3×10^{-8} per nucleotide, which is several orders of magnitude lower than for any other promoter prediction method. The developed method can be used to search for PPSs in various eukaryotic genomes.

Keywords: human genome; promoter; genetic algorithm; multiple alignment

check for updates

Citation: Zaytsev, K.; Fedorov, A.; Korotkov, E. Classification of Promoter Sequences from Human Genome. *Int. J. Mol. Sci.* **2023**, *24*, 12561. https://doi.org/10.3390/ijms241612561

Academic Editor: Irmgard Tegeder

Received: 2 June 2023
Revised: 28 July 2023
Accepted: 3 August 2023
Published: 8 August 2023

1. Introduction

Promoter regions are important for RNA transcription [1]. Such regions are located upstream of DNA coding regions and serve to bind RNA polymerase and several other proteins involved in transcriptional regulation [2]. Eukaryotic promoters can contain highly conserved motifs such as the initiator (Inr), TATA box, DPE, and the bridge, a bipartite core promoter element [3]. These motifs are located in the region −499 to +100, with the +1 position being the transcript start site (TSS). For the positioning of the TSS, sequencing technologies are used that allow for rather precise positioning [4]. The use of OligoCap, CAGE, and deepCAGE makes it possible to find TSSs in in vivo experiments [5,6]. However, the use of experimental methods is still expensive. Therefore, it is easier to use bioinformatics methods for promoter search in sequenced genomes instead [7]. Mathematical methods for promoter search have developed significantly in the last 20 years. Promoter sequence search can greatly improve the accuracy of predicting genes and DNA coding regions [8], as well as accurately locate TSS positions. TSS coordinates are particularly important for genetic engineering and personal medicine because gene expression is regulated by promoter activation and inhibition by various proteins [9,10].

Promoter sequence recognition can be performed using neural networks trained on the existing set of promoters [11], such as the EPD database [12,13], because all promoters in it have been experimentally verified. If the structure and relative positions of individual promoter motifs are known, mathematical tools such as Context Free Grammar can also be used for promoter search [14]. However, promoter sequences are very diverse, which makes their prediction difficult. To overcome this diversity, prokaryotic promoters were divided into several classes based on the sigma factor. Promoter classification allowed us

to reduce the number or false positives for promoter prediction by adjusting the search methods for each class [15,16].

However, all of these methods have a common flaw. They produce a large number of false positives that can interfere with true promoter identification. Even the best of them will make a false positive prediction every few tens of thousands of nucleotides [17–23]. This would result in more than 3×10^5 false positive predictions for the human genome. There are approximately 3×10^4 known promoters in the human genome [12], so such a number of false positives would be an order of magnitude higher than for true promoter sequences. Due to the difficulty in identifying true promoters from the false positives, there is a need for other promoter prediction methods with higher sensitivity for the analysis of complete eukaryotic genomes.

To solve the promoter diversity problem, we used the MAHDS method [24] for the multiple alignment of promoter sequences, which is based on the genetic algorithm and dynamic programming [24]. Its main advantage is the ability to compute a statistically significant multiple alignment for nucleotide sequences that have accumulated a large number of nucleotide substitutions relative to each other. Other methods such as ClustalW [25], Clustal Omega [26], MAFFT [27], T-Coffee [28], and Muscle [29] can only produce statistically significant multiple alignments for sequences with $x < 2.4$, where x is the average number of substitutions per nucleotide [30] among the aligned sequences. This is due to the fact that these methods generate multiple alignments using pairwise alignments in one way or another. When $x > 2.4$, pairwise alignments become statistically insignificant and there is no way to generate statistically significant multiple alignments from them. However, statistically significant multiple alignments can exist for $x > 2.4$ and can be generated for a number of sequences greater than 2. This topic has been discussed in detail in [24].

Unlike other methods, MAHDS is able to generate multiple alignments for sequences with $x < 4.4$ [24,30]. The reason is that MAHDS uses multiple alignment patterns instead of generating multiple alignments from pairwise alignments. For promoter sequences, $x \approx 3.6$ [24]; so, statistically significant multiple alignments cannot be obtained with any method other than MAHDS.

The MAHDS method has previously been used for promoter prediction in the *Arabidopsis thaliana* and *Oryza sativa* genomes [24,31]. The number of false positives was estimated to be $\sim 10^{-8}$ per nucleotide. While promoters have on average $x \approx 3.6$ substitutions per nucleotide, for some of them this number can be much higher. Therefore, even with the MAHDS method, there is a need for promoter classification. In studies [24,31], 500 sequences were randomly selected from the promoter database for a given organism. Multiple alignment was performed on these sequences and a position weight matrix (PWM) was calculated. Such a PWM was used as a class matrix. A promoter class consisted of sequences that had a significant alignment with this PWM. Already-classified sequences were removed from the database and the process was repeated until the class size was less than 100 sequences. This resulted in 25 classes combining 17,787 promoter sequences for the *A. thaliana* genome [24] and 5 classes combining 2458 promoter sequences for the *O. sativa* genome [31]. While this method works, it produces a large number of non-optimized classes. Each set of 500 promoters could contain sequences with $x > 4.4$, which would reduce the promoter class specificity of the PWM. Promoter regions from −499 to +100 were used for PWM generation. However, there is a strong triplet periodicity in the +1 to +100 region [32], which may affect the classification of the promoter sequence and thus the accuracy of promoter prediction.

In this study, we describe a new method for promoter sequence classification based on a genetic algorithm and the MAHDS method [24]. It allowed us to divide the human promoters into four classes, containing more than half of the promoters from the EPD database [12]. Then, using dynamic programming in the same way as in [31], we performed a potential promoter sequence (PPS) search in the human genome. A total of 3,065,317 PPSs were found. We analyzed the PPS intersections with known interspersed repeats, genes, introns, and transposable elements from the human genome. It turns out that there is a

high probability of PPS intersections with all types of ALU sequences, LTR sequences, and transcript start sites (TSSs).

2. Results

2.1. Human Genome Promoter Sequence Classes Creation

We shuffled each of the human promoter sequences (29,598 in total). We then created classes from this set of sequences using an algorithm described in Section 4.2. The maximum size of a class was 121 sequences. Since shuffled sequences should not contain common information, we considered such classes as false positives. In order for the promoter classes to contain no more than 20% false positives by volume, we set the minimum promoter class volume to 605 sequences.

Then, we divided 29,598 human genome promoter sequences into classes. We created four promoter classes that met the minimum size requirement. In total, these four classes contained 17,310 promoter sequences, or 58% of the original set. The volume of each class is shown in Table 1.

Table 1. The table shows the volumes N of the promoter sequence classes created for the human genome. If the volume of a class is less than 605 sequences, the false positive rate would be higher than 20%. Therefore, we stopped at the 4th class of promoter sequences. R is a ratio of the probability of finding an alignment with a promoter sequence from the test sample to the probability of finding an alignment with a promoter sequence from the training sample.

Class	N	R
1	11,780	1.077
2	3139	1.063
3	1624	0.748
4	767	0.648
5	464	0.447
6	161	0.267

For each of the created classes, we calculated R, a measure of class scalability. R is the ratio between the probabilities of finding a promoter sequence in the test sample and in the training sample. The lower R is, the more the class is tuned to the specific sequences from the training set. As we can see, the R value for the first two classes is close to 1, which means that there is no overfitting. As for the other classes, the higher the class number, the less scalable it becomes. This is probably due to the fact that most of the common features of the promoter sequences were already present in the previous classes.

Fasta files and PWMs for the four created promoter classes are available as Supplementary Material.

2.2. PPS Search in Human Genome

We then searched the human genome for the PPS. We searched each chromosome independently for + and − strands. We also searched inverted and shuffled chromosomes for the PPS. Such sequences have the same nucleotide content as the original chromosomes. The PPS search on randomly shuffled chromosomes was performed to estimate the number of false positives for our search method. The PPS search on inverted chromosomes was essential to estimate the number of symmetric PPSs.

In total, we found 3,065,317 non-intersecting PPSs with $Z > 6$ in the human genome. We also found 76,656 sequences with $Z > 6$ on inverted chromosomes. The number of false positives was 186 for the whole genome. The length of the human genome was approximately 3×10^9 nucleotides. The estimation of the number of false positives was performed for shuffled chromosomes for both + and – strands; thus, the number of false positives for the method used was estimated to be $186/6 \times 10^9 \approx 3 \times 10^{-8}$ per nucleotide.

The number of PPSs for human genome chromosomes, shuffled chromosomes, and inverted chromosomes are shown in Table 2.

Table 2. Number of potential promoter sequences (PPSs) found in different chromosomes of the human genome. + and − are DNA strands.

Chr.№	Number of PPSs		Inverted Chromosome		Random	
	+	−	+	−	+	−
1	133,073	133,437	3374	3252	8	8
2	126,853	128,025	3113	3122	8	9
3	94,729	94,889	2218	2182	3	7
4	81,690	82,344	2125	2049	4	6
5	84,473	84,996	1886	1814	5	7
6	84,193	84,668	2308	2310	6	2
7	82,033	82,165	2242	2152	3	4
8	72,840	72,414	1763	1794	7	5
9	63,751	63,647	1511	1573	2	2
10	73,439	73,275	1739	1724	2	4
11	72,425	72,705	1623	1567	2	4
12	72,578	72,840	1952	2008	9	9
13	45,296	45,786	1133	1131	7	3
14	47,289	47,762	1099	1126	2	3
15	47,703	47,644	1024	987	0	3
16	47,886	47,962	1133	1177	4	1
17	52,660	54,031	1194	1289	3	1
18	39,048	39,345	1031	986	3	5
19	40,174	40,246	1603	1592	1	2
20	39,638	39,855	930	939	2	2
21	20,795	20,958	724	711	0	1
22	25,307	25,335	515	507	0	3
X	69,891	71,404	1987	1837	3	8
Y	10,767	10,767	295	305	2	1
Total	3,065,317		76,656		186	

We also calculated Z value distributions for alignments of non-intersecting sequences from the 19th chromosome + strand, the shuffled 19th chromosome + strand, and the *Pr* set with first and second class matrices to select the minimum *Z* value for positive PPS prediction. Such distributions are shown in Figures 1 and 2.

We chose the minimum level Z > 6 for positive PPS prediction as the closest integer to the upper limit of the Z value distribution for alignments of sequences from the shuffled 19th chromosome with the class matrices (Figures 1 and 2). It can be seen that the Z value distribution for alignments of sequences from the 19th chromosome and the Z value distribution for alignments of promoter sequences from the *Pr* set are not offset. Z value distributions for other chromosomes and other class matrices are similar to these, with minor changes.

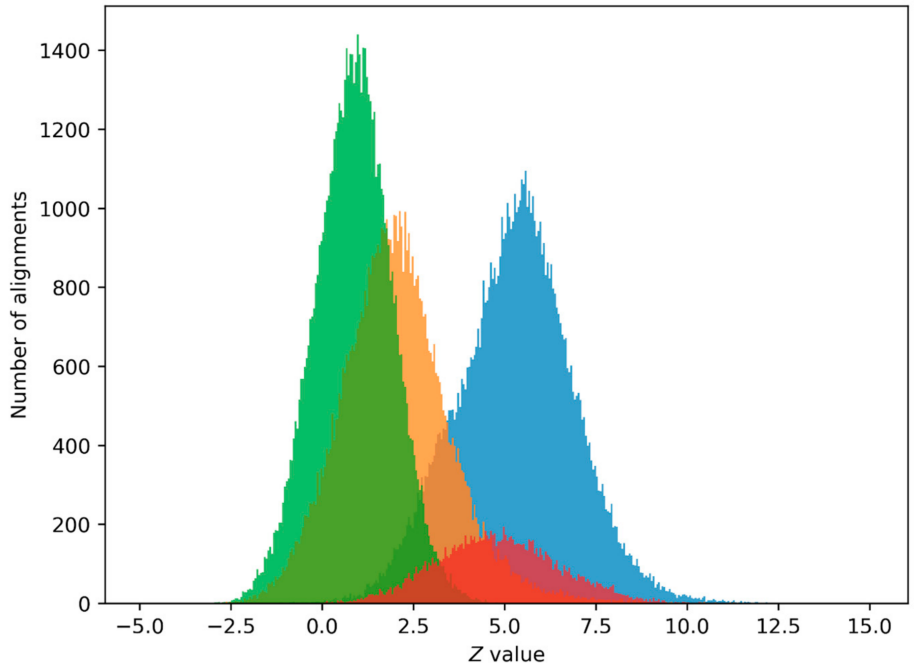

Figure 1. The blue graph is a Z value distribution for alignments of non-intersecting sequences from the 19th chromosome + strand with the first class matrix $W_1(i, j)$ that are at least 490 nucleotides long. The orange graph is a Z value distribution for alignments of sequences from the inverted 19th chromosome + strand, the green graph is a Z value distribution for alignments of sequences from the shuffled 19th chromosome + strand, and the red graph is a Z value distribution for alignments of sequences from the *Pr* set (Section 4.1), which contains 29,598 human promoter sequences.

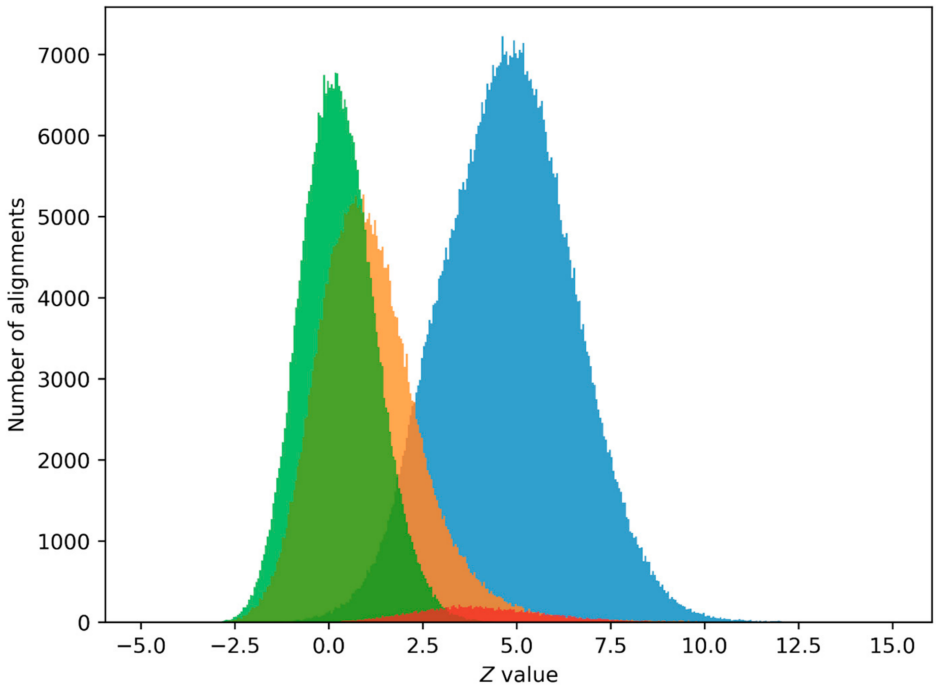

Figure 2. The blue graph is a Z value distribution for alignments of non-intersecting sequences from the 19th chromosome + strand with the second class matrix $W_2(i, j)$ that are at least 490 nucleotides long. The orange graph is a Z value distribution for alignments of sequences from the inverted 19th chromosome + strand, the green graph is a Z value distribution for alignments of sequences from the shuffled 19th chromosome + strand, and the red graph is a Z value distribution for alignments of sequences from the *Pr* set (Section 4.1), which contains 29,598 human promoter sequences.

2.3. PPS Intersection with Annotated Sequences from the Human Genome

We studied the intersection of human genome PPS coordinates with different types of annotated sequences. For PPS intersections with a *j*-type of annotated sequences, the *Zr*(*j*) coefficient represents the deviation from an expected number of intersections with randomly distributed PPSs in the human genome. Near-zero *Zr*(*j*) values are indicative of a random intersection that does not have statistical significance. Large positive values indicate a positive correlation between PPSs and *j*-type sequence positions. Large negative values indicate a negative correlation between the positions of PPSs and *j*-type sequences.

Coordinates of true promoters and promoters for non-coding genes were obtained from the EPD database [12,13,33]. Annotated exon and intron coordinates for the human genome were obtained from the Ensembl database [34,35]. Repeat and transposable element coordinates were obtained from the Dfam database [36,37]. TSS coordinates were obtained from the refTSS v3.3 database [38,39]. This database contains TSS coordinates from several sources, including NCBI Genes, GENCODE, and ENSEMBL.

The results of the PPS intersection with various annotated sequences are shown in Table 3.

Table 3. Number of PPS intersections with different annotated sequences from the human genome. The calculation of *Zr*(*j*) is described in Section 4.7. Promoters ++/−− mean intersection of PPSs with −499:20 promoter regions for genes from the same strand. Promoters +−/−+ mean intersection of PPSs with −499:20 promoter regions for genes from the opposite strand (e.g., PPSs from + strand with promoters from − strand). NC promoters mean the intersection with the −499:20 promoter region for non-coding genes.

j	Annotated Sequences	*Zr*(*j*)	Number of PPSs
1	Promoters ++/−−	57	10,755
2	Promoters +−/−+	−12	4458
3	NC promoters	8	673
4	TSS	58	50,895
5	Exons	39	120,285
6	Introns	60	856,456
7	MIR	−93	66,542
8	AluS	286	196,167
9	AluY	135	38,734
10	AluJ	151	66,919
11	MER	−31	53,258
12	LTR	58	41,789
13	MLT	3	42,846
14	SVA	94	31,567
15	Charlie	−34	6403
16	Tigger	−61	6047
17	MamRep	−16	2885
18	HERV	43	10,963
19	THE	20	12,186
20	L1	−176	111,531
21	L2	−36	70,368
22	L3	−13	4313
23	L4	−22	1313

Table 3. *Cont.*

j	Annotated Sequences	*Zr*(*j*)	Number of PPSs
24	FLAM	21	5604
25	FRAM	12	2526
26	UCON	−12	1502
27	MST	9	7126

The intersection of PPSs with the promoter set has a rather large value of Zr = 57, indicating its high statistical significance. However, the predicted PPSs only intersect with 10,755 true promoters, while it is possible to predict 17,310 true promoters using four PWM matrices for promoter classes. We assume that the reason is that the sites with higher alignment weights were found in close proximity to some of the true promoters. The presence of such sites may indicate the existence of multiple promoters for a gene [40] or alternative reading frames [41]. As a control, we intersected PPSs with a set of true promoters from the alternative strand. Such an intersection has a negative value of Zr = −12, indicating a lack of positive correlation, as it should do.

We also intersected PPSs with a set of promoters for non-coding RNA genes. Such genes can use different RNA polymerases for transcription. This intersection has a rather low positive value of Zr = 8, despite the fact that a large proportion of RNA genes use RNA polymerase II for transcription. We assume that this is due to the difference in use of regulatory elements.

The intersection of PPSs with TSSs has a value of Zr = 58, which is close to that of true promoters. At the same time, we found intersections with only 50,895 out of 223,949 TSSs from the refTSS database [39]. The reason for this is that the promoter classes we created do not include 42% of the promoters for protein-coding genes. Also, some of the TSSs may be related to RNA polymerase III activity, and for such promoters the search method is not effective because PWMs were created for RNA polymerase II promoters.

Among the PPS intersections with different types of repetitive sequences, the highest Zr values were obtained for intersections with Alu elements. This may seem surprising since Alu elements contain the RNA polymerase III promoter [42]. However, we found that more than 75% of the PPSs intersecting with AluS, AluY, or AluJ do not contain their promoter region. Therefore, such high Zr values are not due to the promoter structure, but to the structure of the rest of the Alu elements. Alu elements are relatively rich in CpG residues [42], while about 70% of human proximal promoters contain a CpG island [43]. Alu elements have also been found to host a number of transcription factor binding sites [44] and may contain even more. In several cases, Alu elements have been found to influence the expression of nearby genes [44]. At the same time, Alu elements prefer gene-rich regions of the genome [45] and therefore have a high chance of random prediction. Located within genes, Alu elements have been implicated in alternative splicing, translation regulation, and RNA editing [46].

A large value of Zr = 53 was also obtained for long terminal repeats (LTRs), which have also been found to be involved in gene expression regulation [47]. All other types of repetitive sequences with high positive Zr values are related to either Alu or LTR.

2.4. Comparison with Existing Methods

We compared our method with several existing promoter prediction methods: FPROM [48], TSSW [49], and NNPP [50]. For each method, we generated three sets of sequences. Set_1 contained randomly selected promoter sequences with coordinates −499:20 from the EPD database [33]. Set_2 consisted of random 520 nucleotide sequences with the same nucleotide distribution as the human genome. Set_3 consisted of randomly selected PPSs obtained in this study. Volumes of these sets were denoted as N_s. The results of applying the FPROM, TSSW, and NNPP methods to these sets are shown in Table 4.

Table 4. N_s is the number of sequences in Set_1, Set_2, and Set_3. Set_1 contains randomly selected human promoters from the EPD database. Set_2 consists of randomly shuffled promoter sequences. Set_3 contains randomly selected PPSs from this study.

Method	N_s	Set_1	Set_2	Set_3
FPROM	1000	327	11	24
TSSW	400	219	45	75
NNPP	100	47	42	31
Our method	1000	584	0	1000

Here, positive predictions from Set_1 represent true positives, positive predictions from Set_2 represent false positives, and the number of positive predictions from Set_3 represents the degree of correlation between the studied method and ours.

FPROM is based on a linear discriminant function including functional motifs and oligonucleotide composition description. The method includes two independent functions for TATA and non-TATA promoters. The sensitivity threshold was set at sensitivity = 0.6 for both types of promoters. Here, Ns = 1000 sequences. This method predicts promoters from Set_1 relatively well, as 327 out of 1000 promoter sequences were found. The number of found PPSs from Set_3 was 24, which is significantly more than the 11 false positives found in Set_2.

TSSW recognizes the PolII promoter region and transcription start site. It does not require any parameters. This method gave results comparable to FPROM (Table 4).

NNPP is a method for transcription start site search using a neural network trained on a set of promoters from human and *Drosophila melanogaster* genomes. The minimum promoter score was set to 0.9. NNPP appeared to be incapable of promoter prediction, as the number of true positives and false positives was almost equal (Table 4).

Our method has a higher number of true positives and a lower number of false positives than both FPROM and TSSW, making it more accurate than both. Both FPROM and TSSW make significantly more positive predictions for sequences from Set_3 than for sequences from Set_2. This suggests a positive correlation between these two methods and our method. Sequences predicted by both our method and one of these methods have a much higher chance of being true promoter sequences.

3. Discussion

The previous version of the promoter classification algorithm [24] was started by randomly selecting 500 promoters from the promoter set *Ps* for a given organism. An optimal position weight matrix $W(i, j)$ was then computed using the MAHDS method. Pairwise alignments were generated for each sequence from the *Ps* set with the selected position weight matrix $W(i, j)$, and the Z value was calculated for each of the pairwise alignments. Promoters whose alignments with the position weight matrix $W(i, j)$ had $Z > 5.0$ were grouped into a promoter class and $W(i, j)$ became its class matrix. Then, all of the promoters of the created class were removed from the *Ps* set and the procedure was repeated until the size of the created classes became less than 100 sequences. Such an algorithm was used to classify the promoters from the *A. thaliana* and *O. sativa* genomes [24,31].

Due to the random selection of 500 promoters to be used as seed for each class, such sets could contain sequences with $x > 4.4$. This would result in a suboptimal position weight matrix, as the MAHDS method is only able to generate statistically significant multiple alignments for sequences with $x < 4.4$ [24,30], thus reducing the specificity of the PWM. In this study, we replaced the promoter classification algorithm with a new one based on a genetic algorithm. Here, we created $Ko = 10^2$ seeds for each class. Each seed $Or(k)$ was considered as an organism. The purpose of using the genetic algorithm was to create a seed with the largest possible F_m value, meaning that it would contain the most similar

promoters. To do this, organisms were subjected to crosses, which created a new organism consisting of sequences from two other organisms, and mutations, which replaced several sequences in the organism with sequences from the *Pr* set. The best-performing organism became the seed for the promoter class. In this study, the classes contained promoters with $Z > 6.0$. In this way, the specificity of the $W_1(i, j)$ matrix, which is a promoter class PWM, was significantly improved.

We created four classes of promoter sequences, covering 17,310 of the 29,598 sequences in the EPD database. For each of the classes, we calculated the scalability measure *R*, which shows the ratio between the probability of finding a promoter sequence in the test set and in the training set. The *R* values for the four classes were between 1.077 and 0.648, and for each class, the *R* value was lower than for the previous one. This meant that the position weight matrices for the four promoter classes combined most of the common features of human promoter sequences.

We compared our method with several existing promoter prediction methods: FPROM [48], TSSW [49], and NNPP [50]. Our method showed both a lower number of false positives and a higher number of true positives than any of the methods studied. However, PPSs that are positively predicted via both our method and one of the other promoter prediction methods have an increased chance of being true promoters.

We searched 24 chromosomes from the human genome and found 3,065,317 PPSs, with an estimated number of false positives at 3×10^{-8} per nucleotide, which is several orders of magnitude lower than any other existing promoter search method [51]. We analyzed the PPSs by intersecting their positions with the positions of different types of annotated sequences from the human genome. For each such intersection, we calculated the *Zr* value, which indicated the deviation from an expected number of intersections if the PPSs were randomly distributed in the genome. The intersections of PPSs with true promoters had a value of $Zr = 5$, indicating their high statistical significance. Such an intersection confirms the effectiveness of our promoter prediction method. The intersection with TSSs had a similar value of $Zr = 58$. Those PPSs that intersect with TSSs have a high probability of being true promoters. Thus, an experimental verification of some selected PPSs can be suggested as future work.

The intersection of PPSs with intron sequences had a value of $Zr = 60$, and that with exon sequences was $Zr = 39$. This distribution of PPSs across genes suggests the presence of unknown regulatory pathways related to alternative splicing [52] or alternative reading frames [41]. High *Zr* values for intersections with Alu elements are also evidence for the presence of alternative splicing, referred to as Alu exonization [42]. This phenomenon is widespread, affecting hundreds of human genes. As a result, there may be many more promoter sequences in the human genome than previously thought.

Out of a total of 3,065,317 potential promoter sequences that we found, most of them intersected with known promoters, genes, transposable elements, or dispersed repeats, but 1,241,206 of them were located in non-annotated regions of the human genome. Such PPSs located in non-annotated regions may be associated with unknown copies of known repetitive sequences or transposable elements or unknown genes. They may also indicate the location of microRNAs [53,54]. In this sense, the study of such 1,241,206 PPSs may be of interest for the regulation of the genetic activity of human cells [55].

4. Materials and Methods

4.1. Promoter Sequences

We used 29,598 human genome promoter sequences from the EPD database [33]. We called this set *Pr* and its elements were denoted as $Pr(i)$, $i = 1, 2, \ldots, 29{,}598$. The sequences we used were in the range of −499 to +20, which was chosen to reduce the influence of triplet periodicity [32] that occurs in coding genomic regions on the PWM. We used the GRCh38 release 103 human genome assembly [34].

4.2. Classification of Promoter Sequences from the Human Genome

The task was to include as many promoter sequences from the *Pr* set as possible in a single class. At the same time, there needed to be a minimum number of promoters from other classes in the created class. To solve these tasks and to classify promoter sequences, we developed a mathematical method based on a genetic algorithm.

To speed up the computation, we created a smaller training set Pr_1 containing 10^4 randomly selected promoter sequences from the *Pr* set. We also created a second set of promoter sequences, Pr_2, containing sequences from the *Pr* set that were not selected in the Pr_1 set. This set would be used as a test set to analyze the reproducibility of the promoter search method.

After that, we created an "organism" for the genetic algorithm. It had $Po = 10^2$ randomly selected promoter sequences from Pr_1 set. In this way, we created $Ko = 10^2$ organisms and named them $Or(k)$, $k = 1, 2, \ldots, Ko$. For each organism $Or(k)$, multiple alignment $Al_k(i,j)$ was performed using the MAHDS method at http://victoria.biengi.ac.ru/mahds/auth. Here, i is an index of a promoter in the organism $Or(k)$, $i = 1, 2, \ldots, Po$ and $j = 1, 2, \ldots, L$, where L is the length of the multiple alignment. Columns with a number of bases less than $Po/2$ were deleted from the multiple alignment $Al_k(i,j)$. The result was a transformed multiple alignment $At_k(i,j)$ with a length of L_t. $L_t \leq L$. Then, a frequency matrix $M_k(j,l)$ with $j = 1, \ldots, L_t$ and $l = 1, 2, \ldots, 16$ was calculated for the multiple alignment $At_k(i,j)$.

$$M_k(j, f(At_k(i,j)) + 4(f(At_k(i,j+1)) - 1)) = M_k(j, f(At_k(i,j)) + 4(f(At_k(i,j+1)) - 1)) + 1 \quad (1)$$

When filling the $M_k(j,l)$ matrix, i varies from 1 to Po and j varies from 1 to L_t. $f(a) = 1$, $f(t) = 2$, $f(c) = 3$, and $f(g) = 4$, where a, t, c and g are DNA bases. If $At_k(i, j))$ or $At_k(i, j+1)$ is equal to *, then the $M_k(j,l)$ matrix remains unchanged. Next, we computed the PWM matrix—$V_k(i, j)$—using the $M_k(i, j)$ matrix:

$$V_k(i,j) = \frac{M_k(i,j) - Kp(i,j)}{\sqrt{Kp(i,j)(1 - p(i,j))}} \quad (2)$$

Here, i varies from 1 to L_t, j varies from 1 to 16, $p(i,j) = x(i)y(j)/K^2$, $x(i) = \sum_{j=1,16} M_k(i,j)$, $y(j) = \sum_{i=1,L_t} M_k(i,j)$, and $K = \sum_{i=1,L_t} \sum_{j=1,16} M_k(i,j)$. Then, we transformed the $V_k(i,j)$ matrix so that the R^2 and K_d parameters were the same for all of the matrices with different k values. $R^2 = \sum_{i=1,L_t} \sum_{j=1,16} V_k^2(i,j)$ and $K_d = \sum_{i=1,L_t} \sum_{j=1,16} V_k(i,j)p_1(i)p_2(j)$. We chose $K_d = 0$ and $R^2 = 45{,}000$. Here, $p_1(i)$ is $1/L_t$ for each i; $p_2(k) = p(l)p(m)$, where $p(l)$ and $p(m)$ are the probabilities of l-type or m-type nucleotides in the nucleotide pair (l,m∈{a,t,c,g}): $p(l) = q(l)/L_A$, where $q(l)$ is the number of l-type nucleotides in Al_k, and L_A is the length of all of the sequences in the multiple alignment Al_k. The matrix transformation procedure is described in detail in [56]. The transformed matrix has $R^2 = 45{,}000$ and $K_d = 0$. After the transformation, we obtained the $W_k(i,j)$ matrix. Such a matrix transformation is necessary so that the similarity function F (Section 4.3) has a similar distribution when different matrices $W_k(i,j)$ [56] are aligned with nucleotide sequences.

Next, we created local alignments for each of the promoter sequences from the Pr_1 set with the $W_k(i,j)$ matrix. The alignment algorithm is described in Section 4.3. Then, we estimated the statistical significance of the alignment using the Monte Carlo method and calculated the $Z(k)$ value. This procedure is described in Section 4.4. We calculated the number of promoter sequences $N(k)$ from the Pr_1 set with $Z > Z_0$ and an alignment length greater than 490 nucleotides. The number $N(k)$ is an objective function for an organism $Or(k)$. The calculation of $N(k)$ was performed for each k from 1 to Ko.

Then, we ranked $N(k)$ in descending order and reordered $Or(k)$ accordingly. k varied from 1 to Ko. This meant that $N(1)$ would be the largest of all $N(k)$ and would correspond

to an *Or*(1) organism. For the *Or*(1) organism, we also created local alignments of each of the promoter sequences from the *Pr* set with the $W_1(i, j)$ matrix. For these alignments, the broader promoter region −655:+176 was used to find sequences shifted by up to 30% of their length, relative to the $W_1(i, j)$ matrix. The number of alignments with $Z > Z_0$ and a length greater than 490 nucleotides was called N_f. The same method was used to determine the number of alignments with a broader promoter region of sequences from the Pr_1 and Pr_2 sets. The obtained numbers of alignments with $Z > Z_0$ and a length greater than 490 nucleotides were designated as N_1 and N_2, respectively.

Next, we removed 10 organisms *Or*(*k*), where *k* was in the interval from *Ko*-9 to *Ko*. These were the 10 organisms with the lowest *N*(*k*). Then, we generated 10 new *Or*(*k*). To generate a pair of new organisms, we selected 2 parent organisms with $k = k_1$ and $k = k_2$, $k_1 \neq k_2$. The selection of k_1 and k_2 was performed randomly and the probability of selecting k_1 and k_2 was proportional to the values of $N(k_1)$ and $N(k_2)$.

Two new organisms (children) *Or* were created due to the two-point crossover of the parent organisms. Two coordinates *i* and *j* were randomly chosen in the interval from 1 to *Po*. The first child contained promoter sequences from 1 to $i - 1$ and from $j + 1$ to *Po* of the organism $Or(k_1)$ and promoter sequences from *i* to *j* of organism $Or(k_2)$. The second child contained promoter sequences from 1 to $i - 1$ and from $j + 1$ to *Po* of organism $Or(k_2)$ and promoter sequences from *i* to *j* of organism $Or(k_1)$. Each of the children was subjected to 5 mutations. Each of the mutations could, with a probability of 50%, either swap the order of two promoter sequences in the child or replace one of the promoters with a new one chosen randomly from the Pr_1 set. However, each promoter sequence from the Pr_1 set cannot be present more than once in any organism *Or*, including children. Thus, if a promoter is present more than once in a child, all of its repeats will be replaced by randomly selected sequences from the Pr_1 set.

In total, 5 pairs of parent organisms were selected and 10 children were created. For each of the 10 children, a position weight matrix $W_k(i, j)$ was created and the *N*(*k*) value was calculated. Then, all organisms were ranked in descending order of *N*(*k*), and for the *Or*(1) organism, the N_f, N_1, and N_2 values were calculated. Then, 10 organisms *Or*(*k*) with *k* in the interval from *Ko*-9 to *Ko* were removed and 10 child organisms were created. This process was repeated until N_f did not increase for 60 cycles.

After the genetic algorithm finished, the $W_1(i, j)$ matrix became the promoter sequence class matrix and the N_f promoter sequences were grouped into one class. The dimension of the matrix was $16 \times L_t$. The N_1 and N_2 values were used to calculate the scalability of the promoter search with the $W_1(i, j)$ matrix. The size of the Pr_1 set S_1 was 10^4 sequences and the size of the Pr_2 set S_2 was equal to the size of the *Pr* set minus 10^4. The scalability coefficient $R = \frac{N_1}{N_2}\frac{S_2}{S_1}$ was a ratio of the probability of finding an alignment with a promoter sequence from the Pr_2 set to the probability of finding an alignment with a promoter sequence from the Pr_1 set.

All sequences contained in the created class were removed from the *Pr* set and the algorithm was repeated for the next class. New classes were created until the last class size N_f was smaller than N_0. The value of N_0 was determined by classifying random sequences. The exact value is calculated in Section 2.1.

4.3. Alignment of Promoter Sequences with Position Weight Matrix W(i, j)

S_1 is a promoter sequence of length L_1. We created the sequence S_2, which contained the numbers 1, 2, ..., L_t. L_1 = 520 if the promoter sequence is selected from the *Pr* or Pr_1 sets. If the alignment is made for a wider promoter region from −655 to +176 nucleotides, then L_1 = 832. For the local alignment construction, we calculated the alignment matrix *F*(*i*,*k*) [57]. Here, the S_1 sequence is located along the *X* axis and the S_2 sequence is located along the *Y* axis. To calculate the local alignment, we filled the matrix zero column and

zero row with zeros. $F(0,k) = F(i,0) = 0$, $F_x(0,k) = F_x(i,0) = 0$, $F_y(0,k) = F_y(i,0) = 0$ for each i from 1 to L_1 and k from 0 to L_t. Then, we filled the F matrix first column and first row:

$$F(1,k) = max\begin{cases} F(0,k-1) + B(S_2(k), S_1(1)) \\ F_x(0,k-1) + B(S_2(k), S_1(1)) \\ F_y(0,k-1) + B(S_2(k), S_1(1)) \\ 0 \end{cases} \quad (3)$$

$$F(i,1) = max\begin{cases} F(i-1,0) + B(S_2(1), S_1(i)) \\ F_x(i-1,0) + B(S_2(1), S_1(i)) \\ F_y(i-1,0) + B(S_2(1), S_1(i)) \\ 0 \end{cases} \quad (4)$$

$$F_x(1,k) = 0 \quad (5)$$

$$F_x(i,1) = 0 \quad (6)$$

Here, i varies from 1 to L_1, and k varies from 0 to L_t. Such filling of the first column and the first row is caused by the fact that we need to consider correlations between neighboring nucleotides when creating the alignment. It was impossible to do this for the first row and column, and so they were filled without the $W(i,j)$ matrix and nucleotide correlation. Therefore, we used the $B(i,k)$ matrix instead of the $W(i,j)$ matrix:

$$B(i,k) = 0.25 \sum_{j=1,4}^{4} W(i, j + (k-1) * 4) \quad (7)$$

The other rows of the F matrix were filled, as shown below:

$$F(i,k) = max\begin{cases} F(i-1,k-1) + W(S_2(k), n) & \text{if } t{=}\ 1 \\ F_x(i-1,k-1) + W(S_2(k), n) & \text{if } t{>}\ 1 \\ F_y(i-1,k-1) + B(S_2(k), S_1(i)) & \text{if } t{=}\ 0 \\ 0 \end{cases} \quad (8)$$

$$F_x(i,k) = max\begin{cases} F(i-1,k) - d \\ F_x(i-1,k) - e \end{cases} \quad (9)$$

$$F_x(i,k) = max\begin{cases} F(i-1,k-1) - d \\ F_x(i-1,k-1) - e \end{cases} \quad (10)$$

Here, i varies from 1 to L_1 and k varies from 0 to L_t; $d = 25$; $e = 6$. At the same time, for each (i,k), we calculated $n = S_1(j) + 4(S_1(i) - 1))$ and t. We considered the pairwise correlation of bases when searching for an alignment in the W matrix, and so we introduced the parameter n. To calculate n, we needed to determine the last base added to the alignment before (i,k). If the last base from the S_1 sequence (the one already added to the alignment) was $s(k - t)$, then $j = k -$ t and $n = S_1(k - t) + (S_1(i) - 1) * 4$. If $t = 1$, it corresponded to a diagonal shift in the F matrix and there was no deletion in the S_1 sequence. If $t > 1$, it corresponded to a $t - 1$ base deletion in the S_1 sequence. Here, t was a parameter that characterized the presence or absence of deletions and their length.

Instead of deletions in the S_1 sequence, there may be deletions in the S_2 sequence. In this case, $t = 0$ and the alignment will skip several columns of the $W(i, j)$ matrix. Since the $W(i, j)$ matrix was created for pairs of nucleotides from adjacent columns, its use is incorrect in the case of deletions in the S_2 sequence. Therefore, for such situations, instead of the $W(i, j)$ matrix, the $B(i, k)$ matrix is used, which lacks correlations between pairs of nucleotides.

After filling the $F(i,k)$ matrix, we found the maximum value of its cells and denoted it as F_m and its coordinates as i_m and k_m. Simultaneously with the $F(i,k)$ matrix, we filled the reverse transition matrix [57]. This allowed us to find the local alignment between the S_1

and S_2 sequences. The start and end coordinates of the local alignment were denoted as i_0, k_0, and i_m, k_m, respectively.

4.4. Estimation of the Statistical Significance of a Local Alignment

We estimated the statistical significance of an alignment using the Monte Carlo method. For the Monte Carlo method, we used the $W(i, j)$ matrix, the S_1 and S_2 sequences, the alignment start coordinates i_0 and k_0, the alignment end coordinates i_m and k_m, and the similarity function value F_m. We also created a Q set with 13,000 shuffled S_1 sequences. We then created a global alignment of each sequence from the Q set with the S_2 sequence. To create the global alignment, we removed 0 from the fourth row in Equations (3), (4) and (8). We also changed the initial conditions for F, F_x, and F_y. $F(0,k) = -kd$; $F(i,0) = -id$; $F_x(0,k) = -kd$; $F_x(i,0) = -id$; $F_y(0,k) = -kd$; and $F_y(i,0) = -id$ for each i from 1 to L_1 and k from 0 to L_t. Then, we created a vector $Fr(i) = F_{Q(i)}(i_m,k_m) - F_{Q(i)}(i_0,k_0)$. Here, $F_{Q(i)}$ is a similarity function between S_1 and $Q(i)$ sequences, where i varies from 1 to 13,000. Next, the mean $\overline{Fr}$ of $Fr(i)$ and its variance $D(Fr)$ were determined. Then, the statistical significance was estimated as $Z = (F_m - \overline{Fr})/\sqrt{D(Fr)}$. The use of such a statistical significance estimate Z allowed us to minimize the influence of the alignment length. It also allowed us to compare alignments of sequences with different position weight matrices $W_k(i, j)$.

4.5. Potential Promoter Sequence (PPS) Search in Human Genome

Matrices for classes $W_1(i, j)$ created in Section 4.2 were used for the PPS search in all of the chromosomes of the human genome. The PPS search was performed separately for each chromosome for + and − strands. When searching for PPSs, we chose a window of 700 nucleotides in length and created local alignments of such window with each of the position weight matrices. In this case, the window was the S_1 sequence with length $L_1 = 700$. After creating the alignment, we calculated its statistical significance. But, this time, the number of random sequences in the set was 10^3 and the length of each sequence was 700 nucleotides. Then, the alignment window was moved by 73 nucleotides and all calculations were repeated. We only considered alignments whose length was not less than 490 nucleotides. If the alignment length was less than this, its $Z = 0.0$. Only alignments with $Z > Z_0$ were considered.

A PPS search was performed for each of the created classes. As a result, we obtained vectors $Zp_k(i)$ and $Zm_k(i)$ for each chromosome, where i was the local alignment start coordinate in the chromosome and k was the promoter class number. $Zp_k(i)$ is a vector for the + chromosome strand, $Zm_k(i)$ is a vector for the—chromosome strand. For each vector $Zp_k(i)$ and $Zm_k(i)$, we found the local maximum coordinate i. Here, the local maximum coordinate was i_m, whereby $Zp_k(i_m) > Zp_k(i)$ for each i from $i_m - L_t$ to $i_m + L_t$. We denoted the local maximum i_m vectors as Lp_k and Lm_k for the + and − DNA strands, respectively. Priority was given to alignments with the largest Z value, regardless of the matrix index.

A PPS search was also performed for inverted chromosome sequences *Cpi* and *Cni* (the sequences were written from the end to the beginning) and for shuffled chromosome sequences *Cpr* and *Cnr*. The PPS search within the *Cpr* and *Cnr* sequences allowed us to estimate the number of false positives for the PPS search within the human genome. The calculation results showed that the number of false positives was about 3×10^{-8} per nucleotide.

As a result, we created lists of PPSs for each chromosome. For each PPS found, we show its level of statistical significance Z, the coordinates on the chromosome, the DNA strand (+ or −), the position weight matrix $W_1(i, j)$, and the promoter class represented by this matrix.

4.6. Z_0 Value Selection

We performed a PPS search without a Z_0 threshold on promoter sequences from the *Pr* set and on the shuffled 19th chromosome + strand. We then calculated the number of predictions with different Z values and plotted them on a bar graph (Figure 1). Here, predic-

tions from the *Pr* set represent true positives and predictions from the shuffled chromosome represent false positives. The goal in choosing the Z_0 value was to minimize the number of false positives while keeping the number of true positives as high as possible. $Z_0 = 6$ was chosen as the minimum integer value, providing a near-zero number of false positives.

4.7. Intersection of PPSs with Annotated Sequences from Human Genome

We studied the intersection of the found PPSs with known annotated sequences from the human genome. Such sequences included promoters, exons, introns, transposable elements, tandem and interspersed repeats, and others. We compared the coordinates of the PPSs with the coordinates of the annotated sequences. We considered two sequences to be intersecting if they contained at least 70% of the shorter sequence.

For each annotated sequence type, we also estimated the statistical significance of its intersection with PPSs using the Monte Carlo method. We created 100 sets of random PPS coordinates (*Kor*(*i*), $i = 1, 2, \ldots, 100$) with a distance between them greater than 520 nucleotides. Each set *Kor*(*i*) was intersected with each of the annotated sequence types. Pr(*i*,*j*) is the number of intersections between the *j*-type annotated sequences and the sequence set *Kor*(*i*). We then calculated the mean $\overline{Pr(j)}$ and variance *D*(*Pr*(*j*)). The estimate of statistical significance was calculated as $Zr(j) = (Pr(i,j) - \overline{Pr(j)}) / \sqrt{D(Pr(j)}$. If *Zr*(*j*) > 6.0, then there was a high probability of a positive correlation between the PPSs and the *j*-type annotated sequence locations. If *Zr*(*j*) < −6.0, then there was a high probability that the PPSs would avoid the type *j* annotated sequences. In other cases, there was only a random intersection between the PPSs and certain annotated sequences.

4.8. Comparison with Other Methods for Promoter Prediction

We compared our promoter search method with three different methods for promoter prediction in the human genome. These methods were FPROM, TSSW, and NNPP. Each method was tested on three datasets: Set_1, Set_2, and Set_3. Set_1 contained randomly selected human promoters from the EPD database and the number of positive predictions from Set_1 represents the number of true positives. Set_2 consisted of shuffled promoter sequences. The number of predictions from Set_2 represents the number of false positives. Set_3 contained randomly selected PPSs obtained in this study. The number of positive predictions from Set_3 indicates the degree of agreement with our method.

4.9. Computational Resources

All computations were conducted on a system with 4xIntel Xeon Platinum 8270 (104 cores in total). It took about 2 weeks to create the promoter classes and about 4 weeks to scan all 24 chromosomes.

5. Conclusions

We have developed a new method for the classification of promoter sequences based on a genetic algorithm. This method was applied to the classification of promoters of protein-coding genes from the human genome. For each of the four classes created, we calculated PWMs using the MAHDS method [24]. We then searched for potential promoter sequences on 24 chromosomes of the human genome by performing pairwise alignments of chromosome sections with position weight matrices. We found a total of 3,065,317 PPSs. The number of false positives was estimated to be 3×10^{-8} per nucleotide. This method was compared with three other methods for predicting promoter sequences within the human genome: FPROM [48], TSSW [49], and NNPP [50]. Our method produced the highest number of true positives and the lowest number of false positives of all the methods studied. The number of false positives for our method is several orders of magnitude lower than for any of the existing methods [51].

Supplementary Materials: The supporting information can be downloaded at https://www.mdpi.com/article/10.3390/ijms241612561/s1. The human potential promoter sequence database is available at http://victoria.biengi.ac.ru/cgi-bin/dbPPS/index.cgi. The source code for the project is available at https://github.com/conzaytsev/Potential_promoter_search (accessed on 27 July 2023).

Author Contributions: Conceptualization, E.K.; methodology, K.Z. and E.K.; software, K.Z.; validation, K.Z., A.F. and E.K.; formal analysis, K.Z.; investigation, K.Z.; resources, A.F.; data curation, E.K.; writing—original draft preparation, E.K. and K.Z.; writing—review and editing, K.Z. and E.K.; visualization, K.Z.; supervision, A.F.; project administration, A.F.; funding acquisition, A.F. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partially funded by a grant from the Ministry of Science and Higher Education of the Russian Federation (agreement no. 075-15-2021-1071).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors express their gratitude to Dmitry O. Kostenko for the technical support in this publication.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Smale, S.T.; Kadonaga, J.T. The RNA Polymerase II Core Promoter. *Annu. Rev. Biochem.* **2003**, *72*, 449–479. [CrossRef]
2. Lee, T.I.; Young, R.A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **2013**, *152*, 1237–1251. [CrossRef] [PubMed]
3. Juven-Gershon, T.; Kadonaga, J.T. Regulation of Gene Expression via the Core Promoter and the Basal Transcriptional Machinery. *Dev. Biol.* **2010**, *339*, 225–229. [CrossRef]
4. Lightbody, G.; Haberland, V.; Browne, F.; Taggart, L.; Zheng, H.; Parkes, E.; Blayney, J.K. Review of Applications of High-Throughput Sequencing in Personalized Medicine: Barriers and Facilitators of Future Progress in Research and Clinical Application. *Brief. Bioinform.* **2019**, *20*, 1795–1811. [CrossRef] [PubMed]
5. Georgakilas, G.K.; Perdikopanis, N.; Hatzigeorgiou, A. Solving the Transcription Start Site Identification Problem with ADAPT-CAGE: A Machine Learning Algorithm for the Analysis of CAGE Data. *Sci. Rep.* **2020**, *10*, 877. [CrossRef]
6. Valen, E.; Pascarella, G.; Chalk, A.; Maeda, N.; Kojima, M.; Kawazu, C.; Murata, M.; Nishiyori, H.; Lazarevic, D.; Motti, D.; et al. Genome-Wide Detection and Analysis of Hippocampus Core Promoters Using DeepCAGE. *Genome Res.* **2009**, *19*, 255. [CrossRef]
7. Cassiano, M.H.A.; Silva-Rocha, R. Benchmarking Bacterial Promoter Prediction Tools: Potentialities and Limitations. *mSystems* **2020**, *5*, e00439-20. [CrossRef] [PubMed]
8. Banerjee, S.; Bhandary, P.; Woodhouse, M.; Sen, T.Z.; Wise, R.P.; Andorf, C.M. FINDER: An Automated Software Package to Annotate Eukaryotic Genes from RNA-Seq Data and Associated Protein Sequences. *BMC Bioinform.* **2021**, *22*, 205. [CrossRef]
9. Martin, R.G.; Gillette, W.K.; Rosner, J.L. Promoter Discrimination by the Related Transcriptional Activators MarA and SoxS: Differential Regulation by Differential Binding. *Mol. Microbiol.* **2000**, *35*, 623–634. [CrossRef]
10. Shir-Shapira, H.; Sloutskin, A.; Adato, O.; Ovadia-Shochat, A.; Ideses, D.; Zehavi, Y.; Kassavetis, G.; Kadonaga, J.T.; Unger, R.; Juven-Gershon, T. Identification of Evolutionarily Conserved Downstream Core Promoter Elements Required for the Transcriptional Regulation of Fushi Tarazu Target Genes. *PLoS ONE* **2019**, *14*, e0215695. [CrossRef]
11. Oubounyt, M.; Louadi, Z.; Tayara, H.; To Chong, K. Deepromoter: Robust Promoter Predictor Using Deep Learning. *Front. Genet.* **2019**, *10*, 286. [CrossRef]
12. Périer, R.C.; Junier, T.; Bucher, P. The Eukaryotic Promoter Database EPD. *Nucleic Acids Res.* **1998**, *26*, 353–357. [CrossRef]
13. Dreos, R.R.; Ambrosini, G.; Groux, R.; Cavin Périer, R.; Bucher, P.; Perier, R.C.; Bucher, P. The Eukaryotic Promoter Database in Its 30th Year: Focus on Non-Vertebrate Organisms. *Nucleic Acids Res.* **2017**, *45*, D51–D55. [CrossRef]
14. Datta, S.; Mukhopadhyay, S. A Composite Method Based on Formal Grammar and DNA Structural Features in Detecting Human Polymerase II Promoter Region. *PLoS ONE* **2013**, *8*, e54843. [CrossRef]
15. Amin, R.; Rahman, C.R.; Ahmed, S.; Sifat, M.H.R.; Liton, M.N.K.; Rahman, M.M.; Khan, M.Z.H.; Shatabda, S.; Ahmed, S. IPromoter-BnCNN: A Novel Branched CNN-Based Predictor for Identifying and Classifying Sigma Promoters. *Bioinformatics* **2020**, *36*, 4869–4875. [CrossRef] [PubMed]
16. Shujaat, M.; Wahab, A.; Tayara, H.; Chong, K.T. PcPromoter-CNN: A CNN-Based Prediction and Classification of Promoters. *Genes* **2020**, *11*, 1529. [CrossRef] [PubMed]
17. Solovyev, V.V.; Shahmuradov, I.A.; Salamov, A.A. Identification of Promoter Regions and Regulatory Sites. *Methods Mol. Biol.* **2010**, *674*, 57–83. [CrossRef] [PubMed]
18. de Jong, A.; Pietersma, H.; Cordes, M.; Kuipers, O.P.; Kok, J. PePPER: A Webserver for Prediction of Prokaryote Promoter Elements and Regulons. *BMC Genom.* **2012**, *13*, 299. [CrossRef]

19. Di Salvo, M.; Pinatel, E.; Talà, A.; Fondi, M.; Peano, C.; Alifano, P. G4PromFinder: An Algorithm for Predicting Transcription Promoters in GC-Rich Bacterial Genomes Based on AT-Rich Elements and G-Quadruplex Motifs. *BMC Bioinform.* **2018**, *19*, 36. [CrossRef]
20. Umarov, R.; Kuwahara, H.; Li, Y.; Gao, X.; Solovyev, V.; Hancock, J. Promoter Analysis and Prediction in the Human Genome Using Sequence-Based Deep Learning Models. *Bioinformatics* **2019**, *35*, 2730–2737. [CrossRef]
21. Wang, S.; Cheng, X.; Li, Y.; Wu, M.; Zhao, Y. Image-Based Promoter Prediction: A Promoter Prediction Method Based on Evolutionarily Generated Patterns. *Sci. Rep.* **2018**, *8*, 17695. [CrossRef]
22. De Medeiros Oliveira, M.; Bonadio, I.; Lie De Melo, A.; Mendes Souza, G.; Durham, A.M. TSSFinder-Fast and Accurate Ab Initio Prediction of the Core Promoter in Eukaryotic Genomes. *Brief. Bioinform.* **2021**, *22*, bbab198. [CrossRef] [PubMed]
23. Bondar, E.I.; Troukhan, M.E.; Krutovsky, K.V.; Tatarinova, T.V. Genome-Wide Prediction of Transcription Start Sites in Conifers. *Int. J. Mol. Sci.* **2022**, *23*, 1735. [CrossRef] [PubMed]
24. Korotkov, E.V.; Suvorova, Y.M.; Kostenko, D.O.; Korotkova, M.A. Multiple Alignment of Promoter Sequences from the *Arabidopsis thalianal*. Genome. *Genes* **2021**, *12*, 135. [CrossRef]
25. Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; McGettigan, P.A.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R.; et al. Clustal W and Clustal X Version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948. [CrossRef] [PubMed]
26. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Soding, J.; et al. Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* **2014**, *7*, 539. [CrossRef] [PubMed]
27. Katoh, K.; Frith, M.C. Adding Unaligned Sequences into an Existing Alignment Using MAFFT and LAST. *Bioinformatics* **2012**, *28*, 3144–3146. [CrossRef]
28. Notredame, C.; Higgins, D.G.; Heringa, J. T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *J. Mol. Biol.* **2000**, *302*, 205–217. [CrossRef]
29. Edgar, R.C. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef]
30. Kostenko, D.O.; Korotkov, E.V. Application of the MAHDS Method for Multiple Alignment of Highly Diverged Amino Acid Sequences. *Int. J. Mol. Sci.* **2022**, *23*, 3764. [CrossRef]
31. Korotkov, E.V.; Suvorova, Y.M.; Nezhdanova, A.V.; Gaidukova, S.E.; Yakovleva, I.V.; Kamionskaya, A.M.; Korotkova, M.A. Mathematical Algorithm for Identification of Eukaryotic Promoter Sequences. *Symmetry* **2021**, *13*, 917. [CrossRef]
32. Frenkel, F.E.; Korotkov, E.V. Using Triplet Periodicity of Nucleotide Sequences for Finding Potential Reading Frame Shifts in Genes. *DNA Res.* **2009**, *16*, 105–114. [CrossRef] [PubMed]
33. Eukaryotic Promoter Database. Available online: https://epd.expasy.org/epd/ (accessed on 1 September 2021).
34. Ensembl Genome Browser. Available online: http://ftp.ensembl.org/pub/release-103/fasta/homo_sapiens/dna/ (accessed on 3 March 2021).
35. Howe, K.L.; Achuthan, P.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; Armean, I.M.; Azov, A.G.; Bennett, R.; Bhai, J.; et al. Ensembl 2021. *Nucleic Acids Res.* **2021**, *49*, D884–D891. [CrossRef] [PubMed]
36. The Dfam Community Resource of Transposable Element Families, Sequence Models, and Genome Annotations. Available online: https://www.dfam.org/releases/Dfam_3.3/annotations/ (accessed on 21 April 2021).
37. Storer, J.; Hubley, R.; Rosen, J.; Wheeler, T.J.; Smit, A.F. The Dfam Community Resource of Transposable Element Families, Sequence Models, and Genome Annotations. *Mob. DNA* **2021**, *12*, 2. [CrossRef]
38. A Reference Data Set for Human and Mouse Transcription Start Sites. Available online: http://reftss.clst.riken.jp/datafiles/current/human/ (accessed on 24 May 2022).
39. Abugessaisa, I.; Noguchi, S.; Hasegawa, A.; Kondo, A.; Kawaji, H.; Carninci, P.; Kasukawa, T. RefTSS: A Reference Data Set for Human and Mouse Transcription Start Sites. *J. Mol. Biol.* **2019**, *431*, 2407–2422. [CrossRef]
40. Koenigsberger, C.; Chicca, J.J., 2nd; Amoureux, M.C.; Edelman, G.M.; Jones, F.S. Differential Regulation by Multiple Promoters of the Gene Encoding the Neuron-Restrictive Silencer Factor. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 2291–2296. [CrossRef]
41. Vanderperre, B.; Lucier, J.-F.; Bissonnette, C.; Motard, J.; Tremblay, G.; Vanderperre, S.; Wisztorski, M.; Salzet, M.; Boisvert, F.-M.; Roucou, X. Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. *PLoS ONE* **2013**, *8*, e70698. [CrossRef]
42. Deininger, P. Alu Elements: Know the SINEs. *Genome Biol.* **2011**, *12*, 236. [CrossRef]
43. Deaton, A.M.; Bird, A. CpG Islands and the Regulation of Transcription. *Genes Dev.* **2011**, *25*, 1010–1022. [CrossRef]
44. Polak, P.; Domany, E. Alu Elements Contain Many Binding Sites for Transcription Factors and May Play a Role in Regulation of Developmental Processes. *BMC Genom.* **2006**, *7*, 133. [CrossRef]
45. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial Sequencing and Analysis of the Human Genome. *Nature* **2001**, *409*, 860–921. [CrossRef]
46. Häsler, J.; Strub, K. Alu Elements as Regulators of Gene Expression. *Nucleic Acids Res.* **2006**, *34*, 5491–5497. [CrossRef] [PubMed]
47. Thompson, P.J.; Macfarlan, T.S.; Lorincz, M.C. Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire. *Mol. Cell* **2016**, *62*, 766–776. [CrossRef]
48. Soloviev, V.; Salamov, A. The Gene-Finder Computer Tools for Analysis of Human and Model Organisms Genome Sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1997**, *5*, 294–302.

49. Solovyev, V.V.; Shahmuradov, I.A. PromH: Promoters Identification Using Orthologous Genomic Sequences. *Nucleic Acids Res.* **2003**, *31*, 3540–3545. [CrossRef] [PubMed]
50. Reese, M.G.; MG, R.; Reese, M.G. Application of a Time-Delay Neural Network to Promoter Annotation in the Drosophila Melanogaster Genome. *Comput. Chem.* **2001**, *26*, 51–56. [CrossRef] [PubMed]
51. Umarov, R.K.; Solovyev, V.V. Recognition of Prokaryotic and Eukaryotic Promoters Using Convolutional Deep Learning Neural Networks. *PLoS ONE* **2017**, *12*, e0171410. [CrossRef]
52. Wang, E.T.; Sandberg, R.; Luo, S.; Khrebtukova, I.; Zhang, L.; Mayr, C.; Kingsmore, S.F.; Schroth, G.P.; Burge, C.B. Alternative Isoform Regulation in Human Tissue Transcriptomes. *Nature* **2008**, *456*, 470–476. [CrossRef]
53. Lee, Y.; Kim, M.; Han, J.; Yeom, K.-H.; Lee, S.; Baek, S.H.; Kim, V.N. MicroRNA Genes Are Transcribed by RNA Polymerase II. *EMBO J.* **2004**, *23*, 4051–4060. [CrossRef]
54. Lagos-Quintana, M.; Rauhut, R.; Lendeckel, W.; Tuschl, T. Identification of Novel Genes Coding for Small Expressed RNAs. *Science* **2001**, *294*, 853–858. [CrossRef]
55. Filipowicz, W.; Bhattacharyya, S.N.; Sonenberg, N. Mechanisms of Post-Transcriptional Regulation by MicroRNAs: Are the Answers in Sight? *Nat. Rev. Genet.* **2008**, *9*, 102–114. [CrossRef] [PubMed]
56. Pugacheva, V.; Korotkov, A.; Korotkov, E. Search of Latent Periodicity in Amino Acid Sequences by Means of Genetic Algorithm and Dynamic Programming. *Stat. Appl. Genet. Mol. Biol.* **2016**, *15*, 381–400. [CrossRef] [PubMed]
57. Durbin, R.; Eddy, S.R.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*; Cambridge University Press: Cambridge, UK, 1998; ISBN 0521629713.