Cluster analysis of *S. Cerevisiae* nucleosome binding sites

Y Suvorova¹ and E Korotkov ¹,²

¹ Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences, Moscow, Russia
² National Research Nuclear University MEPhI (Moscow Engineering Physics Institute). Moscow, Russia
E-mail: suvorovay@gmail.com

Abstract. It is well known that major part of a eukaryotic genome is wrapped around histone proteins forming nucleosomes. It was also demonstrated that the DNA sequence itself is playing an important role in the nucleosome positioning process. In this work, a cluster analysis of 67 517 nucleosome binding sites from the *S. Cerevisiae* genome was carried out. The classification method is based on the self-adjusting dinucleotides position weight matrix. As a result, 135 significant clusters were discovered that contain 43225 sequences (which constitutes 64% of the initial set). The meaning of the found classes is discussed, as well as the possibility of the further usage.

1. Introduction

It is known that the length of a DNA molecule in its unfolded form is many times greater than the size of a cell nucleus, where it is located. There exists a complex system of DNA packaging into chromosomes for DNA placement in the nucleus and access to different part of the genome. Nucleosomes formation is one of the first stages of such packaging. The nucleosome consists of the histone proteins complex (two molecules of the H2A, H2B, H3 and H4 proteins), around which nearly 147 base pairs (bp) of DNA are wrapped, constituting 1.67 turns of a double helix. The distance between the adjacent nucleosomes varies from several bases to hundreds of bases. Sequences that are not occupied by nucleosomes are called linkers. In total, up to 80% of a genome could to be covered by nucleosomes. It was demonstrated that the location of nucleosomes in addition to packaging possesses a regulatory function. The location of nucleosomes regulates availability of DNA sequences for transcription factors and other regulated proteins. The part of DNA sequence, which is connected with the histone proteins complex, is called the nucleosome binding site.

Since the DNA molecule is wrapped around the histone, the local DNA rigidity plays an important role. This rigidity is different for different combinations of nucleotides. Therefore, the DNA sequences, where nucleosomes are positioned, are characterized by certain preferences. There are preferred pairs of nucleotides (dinucleotides) in the sites of direct interaction between the DNA and histone proteins. Under the optimum conditions, when forming a nucleosome, the bent DNA should be located uniformly. The strongest contacts of histone proteins with DNA in a nucleosome are separated by 10-11 nucleotides [1]. The hypothesis that the nucleosome positioning is determined by the nucleotide sequence itself was first put forward by E. Trifonov back in 1980 [2]; he also demonstrated that the optimal nucleosome sequences are characterized by the dinucleotides oscillations with the period of ~10. Similar patterns were found practically all studied eukaryotic genomes, including the human [3]. Later, E. Trifonov and his colleagues, as well as the other authors, examined in detail the periodicity of 10 nucleotide pairs and its relationship with various nucleosome sequences [4–7]. However, there are studies that claim that other factors, such as transcription...
factors and motor proteins, which are capable to move the nucleosomes along the DNA strand to a desired position, are in vivo playing the decisive role [8]. Discussions about which of the factors has major influence are still being conducted. Also, some assessments demonstrate that the 10 period appears more likely to be a group property of several nucleosome sequences rather than of a separate binding site [9].

In the past 30 years, a fairly large number of mathematical methods were developed to determine the possible position of a nucleosome along a DNA sequence. Methods of the nucleosome positioning prediction could be conditionally divided into two groups: statistical and biophysical methods [10]. Methods of the first group are based on the statistical properties of a DNA sequence, and consider it as a sequence of symbols; primarily such methods are based on a periodicity of 10 symbols [11,12]. Methods of the second group are based on calculating the flexibility of the double helix section composed of different nucleotides and of the corresponding nucleosome formation energies [13]. Besides, there are methods that combine both these two approaches. A detailed analysis and comparison of the methods could be found in works [10,14].

However, beside the described periodicity of 10 for different dinucleotides, no general properties pertaining to the nucleosome binding sites were identified. The standard methods of bioinformatics were still not able to construct a significant multiple alignment of the given sequences. There are few studies devoted to the nucleosomes multiple alignment using an extremely limited set of data. The authors of the work [15] constructed a multiple alignment for 177 nucleosome binding sites. Therefore, selection of classes and corresponding common properties of the nucleosome binding sites could appear to be the first step in searching for new common properties of these sequences.

2. Methods

2.1. Data preparation

In this paper we are considering the nucleosome binding sites as a symbolic sequence constructed from foursymbol alphabet {“A”, “C”, “T”, “G”}. A total of 67 517 unique nucleosome binding sites in accordance with the experimental map (unredundant map) [16] were extracted from the S. cerevisiae genome. This genome was chosen as the most studied from the point of view of the nucleosomes location. The nucleosome binding sites were having different lengths ranging from 146 to 152 nucleotides. In order to bring all the sequences to the same length, we cut each sequence down to 140 nucleotides preliminary aligning them in the center.

As a result, the set \( N \) of DNA sequences of the length \( L = 140 \) was obtained (\(|N| = 67 517\)). Each sequence was translated into the 16-symbol dinucleotide alphabet {‘AA’, ‘AC’, ‘AT’ ... ‘GG’} by intersecting pairs. And the corresponding sequences of 139 dinucleotides length were received. Alphabet of the 16 symbols size was chosen for two reasons. First, the increase of the size of the alphabet extends the space of the indicators being examined for clustering. Second, from the description of the nucleosome properties it becomes evident that primarily the pairs of nucleotides are playing an important role in the formation of nucleosomes.

In order to determine the level of statistical significance of values calculated in the work, each of the initial DNA sequences of the \( N \) set was randomly shuffled (using the pseudo-random numbers generator). The resulting mixed sequences were also translated into a 16-symbol alphabet and constituted a set of random sequences designated as \( R \).

2.2. Method of classification

The clustering procedure is based on the self-learning weight matrices sized 16 ×139. This is the so-called position weight matrix; the columns of such a matrix correspond to the position in the sequence, and the rows - to one of 16 dinucleotides. At the initialization stage, the matrix is filled with the dinucleotide
frequencies of one sequence \( (S_0) \) chosen from the initial set \( N \). Let us denote the obtained initial matrix as \( M_0 \). Then the matrix is normalized \( (M_0^*) \) using the special procedure, which is described below.

Further, for each dinucleotide sequence \( S_n \) from the set \( N \) a total weight was calculated using the \( M_0^* \) normalized matrix

\[
F_n = \sum_{i=1}^{139} M_0^*[s_i, i]
\]  

(1)

Where \( s_i \) is the \( i \)th element of the sequence \( S_n \), and \( M_0^* \) \([n,i]\) is the element of the \( M_0^* \). In the same way, the weight for each sequence \( R_n \) from the shuffled set was calculated corresponding value: \( F_n = \sum_{i=1}^{139} m[r_i, i] \). The weights of real sequences are normalized \( Z_n = \frac{F_n - \bar{F}}{\sqrt{D(F)}} \). Where \( \bar{F} \) and \( D(F) \) are the mean value and dispersion of the sequences weight calculated on the set of shuffled sequences \( R \). The sequences \( S_n \) with the obtained weight exceeding the specified threshold value: \( Z_n > Z_0^k \) were selected in a new subset \( N_1 \). Based on the sequences from \( N_1 \) a new frequency matrix \( M_1 \) is filled, while the previous one is being deleted. The new matrix undergoes normalization; and the process is repeated. The process includes several iterations \( K \) with a gradual increase in the \( Z_0^k \) threshold.

**Figure 1.** The diagram showing the main algorithm for the nucleosome binding sites classification based on dinucleotide position weight matrix

The set of sequences \( N_k \), which weight that were calculated using the current matrix are above the threshold at the last iteration \( k=K \), are registered as the new class \( N_k \) and are excluded from further consideration. The obtained profile matrix is also registered. Next, a new initial sequence is selected from the remaining \( |N| = |N| - |N_k| \) set of sequences, and a new matrix is being initialized. The process is repeated
from the very beginning. The procedure is repeated until the non-classified sequences remain in the initial \( N \) multiplicity. The diagram of the clustering procedure is presented in Figure 1. The procedure is carried out from different starting points (first sequence \( S_0 \)); and the best result is selected corresponding to the maximum size of the resulting class.

2.3. Matrices normalization procedure
The procedure for normalizing the position weight matrix of dinucleotides \( M \) in a general case is carried out, as follows. For each element of the matrix, the expected frequencies are calculated: 

\[
P_{km} = \frac{(S_k \ast S_m)}{L^2},
\]

where \( S_k \) and \( S_m \) are the corresponding marginal sums \( S_m = \sum_{i=1}^{16} M[i,m] \); \( S_k = \sum_{j=0}^{139} M[k,j] \); and \( L \) is the sum of the matrix \( M \) elements. Afterwards, on the basis thereof mean and dispersion \( V_{ij} = P_{ij} \ast (1 - P_{ij}) \ast L \) are calculated. As a result: 

\[
M^*_{ij} = \left( M_{ij} - M_{ij} \right) / V_{ij}.
\]

2.4. Matrices normalization procedure taking into consideration the triplet periodicity
To normalize the dinucleotides position weight matrix \( M_k \) (16 × 139) considering the triplet periodicity of the sequences a special procedure was created. A triplet matrix \( T \) is filled according to the same nucleotide sequences from the \( N_k \) set, as \( M_k \) at the current iteration \( k \). The matrix \( T_k \) is (4 × 3) frequency matrix. The matrix rows correspond to four types of nucleotides, and the columns - to three codon positions (disjoint triples). Then, the \( T_k \) matrix is transformed according to the following formula: 

\[
T[i,j] = T[i,j] / K_j,
\]

where \( K_j \) is the sum of elements in the \( j \)th column of the \( T_k \) matrix. After that, theoretical probabilities based on the triplets of the \( T_k \) matrix are calculated for each cell \( M[i,j] \):

\[
P_{ij} = P(n_1)/TP_j \ast P(n_2)/TP_j,
\]

where \( P_1/TP \) is the probability of the first nucleotide in the pair in accordance with its position in the triplet in the initial nucleotide sequence, and \( P_2/TP \) is the probability of the second nucleotide. Thus: 

\[
P_{ij} = T[n_1, t_1] \ast T[n_2, t_2].
\]

Then the following expected values are calculated: 

\[
Mean_{ij} = P_{ij} \ast N \quad \text{and} \quad D_{ij} = P_{ij} \ast (1 - P) \ast N.
\]

Here \( N \) is the sum of elements of the \( M \) matrix corresponding column. Then the elements of the \( M \) matrix are transformed according to the following formula

\[
M^*_{ij} = \frac{M_{ij} - Mean_{ij}}{\sqrt{D_{ij}}}.
\]

2.5. Implementation
The software implementation of the described algorithms was created on the C ++ programming language for the Linux operating system and using the MPI library for parallel computations. The calculations were carried out on the Joint Supercomputer Center of the RAS and on the cluster of Institute of Bioengineering RAS. In order to calculate the Fourier transform, the fft library of the R programming language was employed.

3. Results and discussion
Classification analysis of the 67,517 nucleosome binding sites extracted from the \( S. \) cerevisiae genome was carried out with the developed iteration procedure based on the dinucleotides position weight matrices. Analysis of the set of random (shuffled) sequences of the same volume demonstrated that the group size of more than 80 sequences is statistically significant, and corresponds to the 5% significance level. As a result, 135 disjoint significant clusters were found at this level when processing real data. The clusters contained 43225 of 67,517 sequences from the initial set \( N \) (about 64%). The average class size is about 400 sequences.

Many nucleosome binding sites in our study corresponds to protein coding regions. Previously, it was demonstrated that the triplet periodicity is inherent in the most protein coding sequences [17]; and the
triplet periodicity of most known genes belongs to one of 2,5 thousand classes [18]. Thus, the triplet periodicity of different sequences could also affect the results of the classification carried out.

In order to check this hypothesis, an additional classification was carried out, which took into account the influence of the triplet periodicity at the matrix normalization stage (the procedure was described in Section 2.4.). For the new version of the algorithm, the minimum cluster size corresponding to the 5% statistical significance level was also determined on the set of the shuffled sequences $R$. As a result of this new classification 174 significant classes were discovered that contained 39 176 of 67 517 sequences (about 58%).

In order to compare the obtained result with the standard methods, the BLAST (The Basic Local Alignment Search Tool) program was employed to conduct analysis of the nucleosome binding sites in the initial $N$ set [19]. The analysis was aimed at searching for pair similarities of nucleotide sequences within the set. As a result, 3200 cases of the sequences significant similarity inside the set (e-value threshold $<$1e-5) were found. This result means that at the level of the bioinformatics traditional tools there is practically no significant similarity between sequences of the initial set; and it is impossible to identify classes on the basis of such similarity.

As an example of the obtained class features the values of chi-square statistic for each of 139 dinucleotides positions of nucleosome is shown in Figure 2. The chi-square values illustrate the difference between the observed dinucleotide frequencies of the real class and the theoretical frequencies for the position. For clarity’ sake all chi-square values were normalized ($Z(Chi)$). Plots for three sets are shown. The first one is the set of real nucleosome sequences of one of the resulting classes (Real). The volume of the set Real is 325 sequences. The second one is the set of randomly shuffled sequences of the class Real (Shuffled). And the last one is the set of nucleosomes of the same volume as the set Real, randomly chosen from the general set $N$, without the classification procedure (Rand Class). The figure shows that the specificity of the positions of the real class is superior to the rest of the two sets. The plots reveal that the center of nucleosome binding sites is more conservative, this property is noticeable even on randomly selected sequences and is greatly enhanced in the Real class.

![Figure 2](image-url)
Figure 3 demonstrates the Fourier transform of the chi-squared values of the three sets discussed above (see Figure 2). One can see, that only a set of sequences of the Real class has a significant periodicity. The periods 10, 20 that traditionally associated with nucleosome binding sequences. And the most significant period here is 47 symbols. This period has already been described in Trifonov’s work on the study of nucleosomes and it is also associated with the period (10.4) [20].

Figure 3. Fourier transform of the chi-square of the three sets. Real is set of real nucleosome sequences of one of the resulting classes Shuffled – is the set of randomly shuffled sequences of the class Real. Rand Class is the set of nucleosomes of the same volume as the set Real, chosen from the general set N in a random way.

The results obtained in the work demonstrate that there is a fuzzy similarity between the nucleosome binding sequences, which allows us to divide most of such sequences into the statistically significant classes. The obtained position weight matrixes received as the result of the study could be further employed in searching for nucleosome binding sequences using the profile analysis.

Classification procedure, which considers the triplet periodicity, showed that in the case the triplet periodicity has minor influence on the classification result. This could be explained by the fact that the difference in the classes of triplet periodicity is more often observed between genes from different genomes, than within a genome [21]. The result also suggests that there are other factors that play a more significant role in the classification of nucleosome binding sites.

The developed method is not suggesting registration of insertions and deletions inside the nucleosome binding sites, since it is supposed that the structure of a nucleosome itself does not allow such insertions [22]. However, it is known that the experimental methods that determining the coordinates of the nucleosome in a DNA sequence may not be entirely accurate. So, probably, taking into account the shift of the entire section plot to several bases relative to the center would help us to further improve the classification and the obtained position weight matrixes in the future.

Acknowledgements
This work was supported by a grant provided by the Russian Science Foundation (14-24-00175).

References
[1] Xu F and Olson W K 2010 Dna architecture, deformability, and nucleosome positioning J. Biomol. Struct. Dyn.
[2] Trifonov E N 1980 Sequence-dependent deformational anisotropy of chromatin DNA Nucleic Acids Res.
[3] Bettecken T and Trifonov E N 2009 Repertoires of the nucleosome-positioning dinucleotides PLoS One
[4] Radman-Livaja M and Rando O J 2010 Nucleosome positioning: How is it established, and why does it matter? Dev. Biol.
[5] Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore I K, Wang J-P Z and Widom J 2006 A genomic code for nucleosome positioning Nature
[6] Gabdank I, Barash D and Trifonov E N 2009 Nucleosome dna bendability matrix (c. elegans) J. Biomol. Struct. Dyn.
[7] Salih B, Tripathi V and Trifonov E N 2015 Visible periodicity of strong nucleosome DNA sequences J. Biomol. Struct. Dyn.
[8] Hughes A L and Rando O J 2014 Mechanisms Underlying Nucleosome Positioning In Vivo Annu. Rev. Biophys.
[9] Jin H, Rube H T and Song J S 2016 Categorical spectral analysis of periodicity in nucleosomal DNA Nucleic Acids Res.
[10] Xi L, Mittendorf Y F, Xia L, Flatow J, Widom J and Wang J P 2010 Predicting nucleosome positioning using a duration Hidden Markov Model BMC Bioinformatics
[11] Xing Y Q, Liu G Q, Zhao X J and Cai L 2013 An analysis and prediction of nucleosome positioning based on information content Chromosom. Res
[12] Morozov A V., Fortney K, Gaykalova D A, Studitsky V M, Widom J and Siggia E D 2009 Using DNA mechanics to predict in vitro nucleosome positions and formation energies Nucleic Acids Res.
[13] Ioshikhes I, Bolshoy A, Derenshteyn K, Borodovsky M and Trifonov E N 1996 Nucleosome DNA Sequence Pattern Revealed by Multiple Alignment of Experimentally Mapped Sequences J. Mol. Biol.