Patterns in nucleosomal DNA sequences: tools and examples.
Supplementary Appendix S1 to ‘Patterns in nucleosomal DNA as packing or regulatory’.

Contents

1 Background 3

2 Workflow 3
   2.1 Distribution of frequency of occurrence of dinucleotides along nucleosome sequences 5
   2.2 Cleavage site 5
   2.3 Determination of nucleosome position using dyad-symmetry of dinucleotide frequency profiles 6
   2.4 Patterns of dinucleotide frequency distributions and their periodograms 7

3 Dinucleotide shuffling 12

4 Software implementation of dnpatterntools v1.0 12
   4.1 Interface to Galaxy 12

5 Figures 13
List of Figures

A Example of transformation preceding a computation of a positional frequency profile of dinucleotide occurrences in all nucleosome sequences. The fasta sequence is transformed into a binary sequence of zeros and ones, in which ones indicate a presence of dinucleotide in that position. In this example the dinucleotide is AA. First column is a binary representation, second column is original sequence and third column shows a number of dinucleotide occurrences in the sequence. 14

B Gaussian smoothed coverage profile with window=70 on positive (+) and negative (-) strands. Identified peak maximums are shown with the red lines. ................................................................. 14

C A. Frequency profiles of selected dinucleotides on forward and complementary sequences of nucleosomes in mouse. Cleavage site is visibly expressed in profiles of all cell types. B. Pearson correlation coefficient between the profiles on forward and reverse complement sequences computed within sliding window of 146bp at the beginning positions of dinucleotide frequency profiles in A. ......................... 15

D AA/TT patterns and periodograms in human and mouse. .............. 16

E CC/GG patterns and periodograms in human and mouse. .............. 17

F WW/SS patterns and periodograms in human and mouse. .............. 18

G WW pattern and periodogram in human and mouse. .................... 19

H SS pattern and periodogram in human and mouse. ...................... 20

I RR/YY patterns and periodograms in human and mouse. ............... 21

J Example of dinucleotide frequency profiles and their correlations in shuffled nucleosome sequences. Panel A. Frequency profiles computed from a batch of full length (219bp) mouse nucleosome sequences shuffled by ushuffle tool. Panel B. Correlation profile. ....................... 22
1 Background

The software tools and protocols to determine nucleosome occupancy and positioning from the micrococcal nuclease sequencing (MNase-seq) data are established already [8, 36]. A nucleosome's occupancy and positioning information are obtained from profiles of peaks resulting from the MNase-seq data alignment to a reference genome of an organism under investigation. Peak positions usually are determined using a traditional gaussian [22] or improved wavelet smoothing [24]. In addition to positioning information (a) shift and occupancy change events for nucleosomes can be computed if data for cases and controls is available [8].

However, practical software tools to compute patterns of dinucleotide frequency distributions from nucleosome sequences have not been sufficiently described. We attempted to describe in more detail an algorithm used in computation of previously characterized patterns in nucleosomal DNA [12, 14]. We describe dnpatterntools v1.0 software utilities to reproduce patterns already characterized and to compute new patterns.

2 Workflow

A workflow to compute patterns of dinucleotide frequency distributions from nucleosome sequences consists of several steps:
1. computation of distribution of frequency of dinucleotide occurrences in a bulk of aligned sequences; 

2. determination of nucleosome’s position in the sequences; 

3. selection and symmetrization of dinucleotide frequency profiles from the determined interval; 

4. computation of frequency profiles of composite dinucleotides WW/SS (W = A or T and S=C or G) and RR/YY (R=A or G and Y=C or T); 

5. application of normalization and smoothing to the frequency profiles to remove noise and to compute periodograms.

To illustrate this workflow practically we use three datasets of nucleosome sequences: from human CD4+ cells [25], from apoptotic lymphocyte cells [4] and from nucleus accumbens cells of mouse brain [30].

The sequences from the CD4+ and apoptotic lymphocyte cells were characterized previously [12]. The nucleosome sequences from mouse nucleus accumbens brain cells [30] were obtained from NCBI GEO archive under accession GSE54263. The Fastq files with the accessions SRR1138261, SRR1138262, SRR1138263 representing MNase-seq 99bp length reads of three biological replicates from control mice were aligned to mm9 reference genome by BWA [17]. The BAM files were converted into BED and genome coverage was computed by BEDTools [23]. The profile of peaks from genome coverage files was generated by Gaussian smoothing using 70bp window size [32]. Genome-wide peak summits exceeding value of 10 were discarded. An example of smoothed coverage signal (shaded areas) and summits (red lines) is presented in Figure B. The summits, positions of which were closest to gene transcription start sites were selected. The reads overlapping these summits were extended by 20bp in 5’ direction and 100 bp in 3’ direction [7] resulting in the 219 bp long sequence (99 bp read plus 120 bp extensions). The DNA sequences in these regions were extracted from mouse mm9 reference by BEDTools, aligned by the experimental 5’-end and formatted into Fasta file.
2.1 Distribution of frequency of occurrence of dinucleotides along nucleosome sequences

In the given nucleosome sequences at each position an occurrence of dinucleotide is coded by 1. All other dinucleotides (positions) are 0 as illustrated by Figure A. Frequency of occurrence of dinucleotide at a single position along the sequence is simply a sum of occurrences of that dinucleotide normalized by the number of sequences. The diprofile program (see Table C) computes frequency distribution of a single dinucleotide along sequences in Fasta file. The dnp-subset-dinuc-profile.sh script calls diprofile and computes frequency distribution of all dinucleotides along sequences in Fasta file in a forward original and in a complementary representation of each sequence. Patterns of dinucleotide frequency distributions represent statistical sequence-specific features of nucleosomal DNA. Patterns originating from different organisms, conditions or experimental manipulations may have signatures characteristic only to that particular condition [14].

2.2 Cleavage site

Nucleosomal DNAs are generally obtained from the purified chromatin stabilized with formaldehyde and digested with MNnase which cleaves sequence specific linker sites [24]. However, sequence fragments resulting from MNase digestion have substantial variability from 10 to 20bp in the precise fragment ends [9]. In dinucleotide frequency profiles computed from nucleosome sequences obtained by MNase-seq and aligned by experimental end a cleavage site manifests as a narrow large peak because of the sequence specificity. In both, profiles of dinucleotide frequencies of mouse cells and of human CD4+ and apoptotic lymphocyte cells the region of a cleavage site can be identified by a large peak at the beginning of the computed frequency profiles. A cleavage site peak is a positive indicator that the batch of sequences obtained from MNase-seq data contains sequences of nucleosomes. Panel A of Figure C shows the large peak at a cleavage site region present in the beginning of the frequency profiles in all datasets - mouse and human.
2.3 Determination of nucleosome position using dyad-symmetry of dinucleotide frequency profiles

Dyad-symmetry is a hallmark of the nucleosome DNA sequence [20]. The peak arrangements in patterns of dinucleotide (most often AA, TT, AT, CC, GC or GG) frequency distribution along the nucleosomal sequence have a recognizable dyad-symmetry. These dinucleotides statistically are preferred [11] and areperiodically distributed along nucleosome DNA sequence [20]. These dinucleotide preferences were investigated and reported by studies in vitro [19],[2], statistically [6, 9], from analysis of nucleosome stability [11] and computationally [3, 5, 14]. Dyad-symmetry feature helps to determine a position of a nucleosome in a bulk of sequences aligned by experimental end - because at the nucleosome position centered on the dyad the forward and complementary profiles of dinucleotides will have a maximum positive correlation. Panel B in Figure C shows Pearson correlation coefficient at each position along the sequence computed between forward (fw) and reversed complement (rc) of frequency profiles for selected dinucleotides within the window corresponding to the nucleosome size of 146bp. In such obtained matrix of Pearson correlation coefficients a maximum positive correlation between of fw and rc frequency profiles of either AA, TT, TA, CC, GG and GC dinucleotides or combination will indicate a nucleosome position - same for all sequences in a batch.

A challenge is that identification of nucleosome position via correlation cannot be fully automated, because correlations vary in both: along the sequence and in different conditions. The position in which fw and rc frequency profiles of one or several dinucleotides attain maximum positive correlation has to be at a close proximity to the cleavage site which is indicator of a nucleosome start. In mouse strongest correlations were found between fw and rc frequency profiles for AA/TG dinucleotides. However, in human cells the strongest correlations were found for AT and GC dinucleotides. Panel B of Figure C shows correlation profiles obtained for all three cases: nucleosomes in mice brain, human CD4+ cells and apoptotic lymphocyte cells. Solid dark red line represents inferred a most likely start position of
the nucleosome from the dinucleotide frequency profiles. For mouse it is at a position 25 from the beginning of the sequences aligned by the experimental end; for human this position is 105 and for apoptotic lymphocyte cells it is 109. The correlation profiles for apoptotic lymphocyte cells is a good illustration that the cleavage site is to be used as a guide, because strong positive correlations are observed also in regions in which a presence of nucleosome is unlikely.

2.4 Patterns of dinucleotide frequency distributions and their periodograms

Nucleosome sequences in yeast are characterized by a very clear pattern of AA/TT frequency distribution with peaks occur each 10 base pairs \[14, 29\]. In other organisms dinucleotides other than AA may elicit clearer patterns of frequency distributions. For example, in human, mouse and flies the GC/CG/CC/GG dinucleotide periodicity correlates better with nucleosome positioning \[18, 25, 33\]. It was also shown that composite patterns of RR/YY (purine-purine / pyrimidine-pyrimidine) can be associated with nucleosome stability \[14\]. In addition to frequency distribution of all 16 dinucleotides we separately compute distribution of frequencies of composite dinucleotides strong-weak SS/WW (S = C or G, W= A or T) and purine/pyrimidine RR/YY (R = A or G, Y = C or T). This is done by the \texttt{dnp-compute-composite.sh} utility (Table C).

The frequency distributions on original forward sequences and their complement should be equally represented in the pattern. Therefore, for each dinucleotide its fw and rc patterns are averaged at each position. This step is called symmetrization. The symmetrized patterns of dinucleotide frequency distributions are computed by utility \texttt{dnp-symmetrize.sh}. Finally, to improve a representation of the patterns they are smoothed by applying a moving average filter. Usually, a several positions of a pattern are trimmed from both ends, since most often they represent noise. Smoothing is accomplished with the program \textit{Fourier_Transform} that is wrapped into \texttt{dnp-smooth.sh} utility, which also applies trimming. Our experience suggests
that the best size for moving average filter is 3 positions and optimal trimming is 4 positions from each end.

A spectral decomposition of a pattern reveals the strongest periodical components that make the pattern. The dinucleotide frequency distributions in nucleosome sequences are expected to have peaks at 10bp and periods multiple of 10. Spectral decomposition may serve as means to identify a leading dinucleotide pattern in each condition. Table A summarizes periods of peaks around 10bp across conditions and major patterns in the data used in this workflow, illustrating variability in peak periods around 10 base pairs. By performing random sampling with replacement. From nucleosome fasta sequences of apoptotic lymphocytes (n=711873), CD4+ cells (n=58150750) and control mice NAC (n=210000) the 50% of sequences 31 times were randomly sampled by seqtk sample function. For each sample of the sequences the dnpatterntools pattern computation pipeline was applied and pattern periodograms of the dinucleotides and conditions were computed. In the [9.5, 11] bp interval of the periodogram a period of a maximal peak was identified. From a sample of 31 period of maximal peak a mean period and its standard error were estimated. These estimates are shown in Table A. ANOVA analysis of the sampled (10bp vicinity) period means in different dinucleotides and conditions did not reveal significant differences between the dinucleotides or conditions. Scripts and results of this estimation procedure can be found on dnpatterntools github site https://github.com/erinijapranckeviciene/dnpatterntools under periods.

In Section 5 the patterns and periodograms are shown for the dinucleotides of interest AA/TT, CC/GG, WW/SS and RR/YY in Figures D, E, F and I. Because of the substantial differences in scale, the WW and SS plots are also presented separately in Figures G and H. The AA/TT, CC/GG, WW/SS and RR/YY patterns of human CD4+ and apoptotic lymphocyte cells reported in [12] are identical to the patterns reproduced in this study. This result ensures that patterns from the nucleus accumbens brain cells of control mice [30] can be compared with the patterns previously obtained, since they were computed by the same pipeline, which generated a reproducible result.
Table A: Variability of 10bp vicinity period in periodograms of dinucleotide frequency distribution patterns in human CD4+ and apoptotic and control mouse NAC cells.

The plotting commands and data are available on github in the same repository as dnpatterntools.

To identify patterns that have a most pronounced spectral power at 10bp range we used above described sampling data. The 10bp period integral spectral power was calculated by summing all components from [10.1 bp, 10.9 bp] interval of the periodogram. The dnpatterntools fourier function computes periodogram that is detrended and normalized which makes integral spectral power comparable across different periodograms. Samples of integral spectral power at 10bp period were computed for the patterns of each AA, TT, CC, GG, SS, WW, RR and YY dinucleotide in each apoptotic lymphocytes, human CD4+ cells and control mouse NAC 31 times from 50% of sequences each time sampled from the original fasta (apoptotic lymphocytes (n=711873), CD4+ cells (n=58150750) and control mice NAC (n=210000) ) using seqtk sample with replacement. For each of these dinucleotides and conditions 31 instances of integral spectral power at 10bp period was computed.

The means of the spectral power between the dinucleotides in each condition were compared using ANOVA test implemented in R. To estimate which differences are significant in multiple hypothesis comparison we used TukeyHSD function im-
plemented in R. Table B shows ranking of the patterns by a magnitude of the mean of the integral spectral power at 10bp period, its standard deviation and standard error for each dinucleotide and condition. The patterns for which the spectral power at 10bp period was significantly different from majority of other patterns (p.adjusted < 0.05) in multiple comparisons are indicated by * in Table B. Results of multiple hypothesis comparison and plots produced by TukeyHSD function supporting Table B are available on dnpatterntools github site under periods.
<table>
<thead>
<tr>
<th>Dinucleotide</th>
<th>Mean spectral power at 10bp period ± sd (stderr)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Apoptotic lymphocytes</td>
</tr>
<tr>
<td>RR*</td>
<td>0.1037 ± 0.0286(0.0051)</td>
</tr>
<tr>
<td>YY*</td>
<td>0.1010 ± 0.0254(0.0046)</td>
</tr>
<tr>
<td>WW</td>
<td>0.0992 ± 0.0258(0.0046)</td>
</tr>
<tr>
<td>SS</td>
<td>0.0975 ± 0.0311(0.0056)</td>
</tr>
<tr>
<td>CC</td>
<td>0.0810 ± 0.0210(0.0038)</td>
</tr>
<tr>
<td>GG</td>
<td>0.0808 ± 0.0222(0.0040)</td>
</tr>
<tr>
<td>AA</td>
<td>0.0805 ± 0.0181(0.0033)</td>
</tr>
<tr>
<td>TT</td>
<td>0.0761 ± 0.0174(0.0031)</td>
</tr>
<tr>
<td></td>
<td>CD4+ cells</td>
</tr>
<tr>
<td>RR*</td>
<td>0.1113 ± 0.0303(0.0054)</td>
</tr>
<tr>
<td>YY*</td>
<td>0.1096 ± 0.0264(0.0047)</td>
</tr>
<tr>
<td>SS*</td>
<td>0.0988 ± 0.0344(0.0062)</td>
</tr>
<tr>
<td>WW</td>
<td>0.0915 ± 0.0219(0.0039)</td>
</tr>
<tr>
<td>GG</td>
<td>0.0767 ± 0.0179(0.0032)</td>
</tr>
<tr>
<td>AA</td>
<td>0.0758 ± 0.0191(0.0034)</td>
</tr>
<tr>
<td>CC</td>
<td>0.0735 ± 0.0190(0.0034)</td>
</tr>
<tr>
<td></td>
<td>Control mouse nucleus accumbens cells</td>
</tr>
<tr>
<td>YY*</td>
<td>0.1082 ± 0.0276(0.0050)</td>
</tr>
<tr>
<td>RR*</td>
<td>0.1036 ± 0.0293(0.0053)</td>
</tr>
<tr>
<td>SS</td>
<td>0.0992 ± 0.0309(0.0055)</td>
</tr>
<tr>
<td>WW</td>
<td>0.0989 ± 0.0338(0.0061)</td>
</tr>
<tr>
<td>AA</td>
<td>0.0812 ± 0.0232(0.0042)</td>
</tr>
<tr>
<td>TT</td>
<td>0.0805 ± 0.0216(0.0039)</td>
</tr>
<tr>
<td>CC</td>
<td>0.0776 ± 0.0147(0.0026)</td>
</tr>
<tr>
<td>GG</td>
<td>0.0775 ± 0.0166(0.0030)</td>
</tr>
</tbody>
</table>

Table B: Ranking of the patterns by 10bp spectral power in each condition.
3  Dinucleotide shuffling

Random shuffling of nucleotides in sequences is a technique that creates randomness and destroys a periodical structure of dinucleotide occurrences in nucleosome sequences. To make our tools complete we included as extra tool a wrapper to the uShuffle program [16] in the utility shuffle_sequences.sh. Figure J illustrates application of shuffling to mouse nucleosome sequences. Panel A shows dinucleotide frequency distributions in shuffled nucleosome sequences of mouse. Panel B depicts correlation profiles computed from shuffled profiles. Obviously, randomized sequences do not have structure that was observed in original sequences (see panel A in Figure C). Similarly, the correlation profiles do not have any prominently expressed correlation value that could be associated with the likely nucleosome position compared to the real data as in panel B in Figure C.

4  Software implementation of dnpatterntools v1.0

The dnpatterntools consist of core programs and tool utilities. The core programs are written in C++ using a SeqAn library [10]. The SeqAn library is a collection of C++ header definitions of functions specifically written to work with genomic data (Fasta, BAM and VCF). A C++ program that uses SeqAn library depends only on the SeqAn function headers. The tool utilities in dnpatterntools are written in shell as commands that integrate also into Galaxy environment [1]. Each has a corresponding fully functional Galaxy xml wrapper. Table C summarizes core programs and tool utilities. Input and output and functionality of each utility is demonstrated by a test script which is a part of the package. The dnpatterntools are available in github repository https://github.com/erinijapranckeviciene/dnpatterntools.

4.1  Interface to Galaxy

All tool utilities in dnpatterntools have fully functional interface to Galaxy [1]. A Planemo configured Galaxy server can be launched calling planemo serve from within the tools directory on a local Linux machine. The Galaxy workflow of dnpatterntools
## Core utilities

<table>
<thead>
<tr>
<th>C++</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>binstrings</code></td>
<td>Binary strings from sequences, more in Figure A.</td>
</tr>
<tr>
<td><code>diprofile</code></td>
<td>Distribution of dinucleotide frequency in sequences.</td>
</tr>
<tr>
<td><code>corrprofile</code></td>
<td>Correlation between forward and reverse complement profiles.</td>
</tr>
<tr>
<td><code>Fourier_Transform</code></td>
<td>Computes periodogram, normalization and smoothing [12].</td>
</tr>
</tbody>
</table>

## Tool utilities

<table>
<thead>
<tr>
<th>Shell</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>dnp-binary-strings.sh</code></td>
<td>Binary strings for multiple dinucleotides.</td>
</tr>
<tr>
<td><code>dnp-compute-composite.sh</code></td>
<td>Composite WW/SS and RR/YY dinucleotide profiles.</td>
</tr>
<tr>
<td><code>dnp-correlation-between_profiles.sh</code></td>
<td>Correlations for all dinucleotides.</td>
</tr>
<tr>
<td><code>dnp-fourier-transform.sh</code></td>
<td>Periodogram for all dinucleotides.</td>
</tr>
<tr>
<td><code>plot-selected.sh</code></td>
<td>Gnuplot of selected columns.</td>
</tr>
<tr>
<td><code>dnp-select-range.sh</code></td>
<td>Select profiles within interval.</td>
</tr>
<tr>
<td><code>dnp-smooth.sh</code></td>
<td>Smoothing by moving average.</td>
</tr>
<tr>
<td><code>dnp-subset-dinuc_profile.sh</code></td>
<td>Frequency profiles of all dinucleotides.</td>
</tr>
<tr>
<td><code>dnp-symmetrize.sh</code></td>
<td>Symmetrization of frequency profiles.</td>
</tr>
</tbody>
</table>

Table C: List of utilities in dnpatterntools

Utilities is described in the document galaxy-workflow.pdf in the dnpatterntools github repository.

## 5 Figures
Figure A: Example of transformation preceding a computation of a positional frequency profile of dinucleotide occurrences in all nucleosome sequences. The fasta sequence is transformed into a binary sequence of zeros and ones, in which ones indicate a presence of dinucleotide in that position. In this example the dinucleotide is AA. First column is a binary representation, second column is original sequence and third column shows a number of dinucleotide occurrences in the sequence.

Figure B: Gaussian smoothed coverage profile with window=70 on positive (+) and negative (-) strands. Identified peak maximums are shown with the red lines.
Figure C: A. Frequency profiles of selected dinucleotides on forward and complementary sequences of nucleosomes in mouse. Cleavage site is visibly expressed in profiles of all cell types. B. Pearson correlation coefficient between the profiles on forward and reverse complement sequences computed within sliding window of 146bp at the beginning positions of dinucleotide frequency profiles in A.
Figure D: AA/TT patterns and periodograms in human and mouse.
Figure E: CC/GG patterns and periodograms in human and mouse.
Figure F: WW/SS patterns and periodograms in human and mouse.
Figure G: WW pattern and periodogram in human and mouse.
Figure H: SS pattern and periodogram in human and mouse.
Figure I: RR/YY patterns and periodograms in human and mouse.
Figure J: Example of dinucleotide frequency profiles and their correlations in shuffled nucleosome sequences. Panel A. Frequency profiles computed from a batch of full length (219bp) mouse nucleosome sequences shuffled by ushuffle tool. Panel B. Correlation profile.
References


[34] Zhang, Y. et al. (2008) Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. BMC Genomics, 9, 537