The roles of the monomer length and nucleotide context of plant tandem repeats in nucleosome positioning

Victor G. Levitskya, Vladimir N. Babenkb and Alexander V. Vershininc*

aLaboratory of Molecular Genetics Systems, Institute of Cytology and Genetics, Novosibirsk, 630090, Russia; bLaboratory of Human Molecular Genetics, Institute of Cytology and Genetics, Novosibirsk, 630090, Russia; cLaboratory of Molecular Genetics, Institute of Molecular and Cellular Biology, Novosibirsk, 630090, Russia

Communicated by Anna Panchenko

(Received 20 February 2012; final version received 7 November 2012)

Similar to regularly spaced nucleosomes in chromatin, long tandem DNA arrays are composed of regularly alternating monomers that have almost identical primary DNA structures. Such a similarity in the structural organization makes these arrays especially interesting for studying the role of intrinsic DNA preferences in nucleosome positioning. We have studied the nucleosome formation potential of DNA tandem repeat families with different monomer lengths (ML). In total, 165 plant tandem repeat families from the PlantSat database (http://w3lamc.umbr.cas.cz/PlantSat/) were divided into two classes based on the number of nucleosome repeats in one DNA monomer. For predicting nucleosome formation potential, we developed the Phase method, which combines the advantages of multiple bioinformatics models. The Phase method was able to distinguish interfamily differences and intrafamily monomer variation and identify the influence of nucleotide context on nucleosome formation potential. Three main types of nucleosome arrangement in DNA tandem repeat arrays – regular, partially regular (partial), and flexible – were distinguished among a great variety of Phase profiles. The regular type, in which all nucleosomes of the monomer array are positioned in a context-dependent manner, is the most representative type of the class 1 families, with ML equal to or a multiple of the nucleosome repeat length (NRL). In the partially regular type, nucleotide context influences the positioning of only a subset of nucleosomes. The influence of the nucleotide context on nucleosome positioning has the least effect in the flexible type, which contains the greatest number of families (65). The majority of these families belong to class 2 and have nonmultiple ML to NRL ratios.

Keywords: tandem DNA repeat families; nucleosome formation potential; Phase method; nucleosome arrangement; nucleotide context

Introduction

Genomes are packaged into nucleosomes, which constitute the chromatin of most eukaryotic cells. The chromatin structure is based on repeated units in which nucleosome core particles, each containing an octamer of histone proteins, are enwrapped by approximately 146 bp of DNA, which is sharply bent and tightly wrapped in 1.65 superhelical turns (Luger, Mader, Richmond, Sargent, & Richmond, 1997). The linker DNA, which is associated with histone H1, connects the core particles. Such repeated units, or nucleosome repeats, are regularly spaced along the DNA, resembling “beads-on-a-string” and representing the building blocks of chromatin. The length of DNA in the nucleosome repeat unit varies between organisms and even between different cell types in the same organism (Kornberg, 1977; van Holde, 1989). Nucleosomes are positioned if their individual locations on the DNA sequence are fixed (van Holde, 1989). The role of DNA sequences in nucleosome formation has been discussed since the early 1980s (Trifonov & Sussman, 1980). Subsequent analysis has shown that a (10.4 ± 0.2)-base periodicity of certain dinucleotides is a ubiquitous hallmark of nucleosome positioning (Drew & Travers, 1985; Ioshikhes, Bolshoy, Derenshteyn, Bordo, & Trifonov, 1996; Thastrom et al., 1999). DNA sequences with high GC content or with AA or TT dinucleotides appearing in periodic intervals favor nucleosome formation, whereas sequences with poly(dA/dT) tracts disfavor nucleosome deposition (Field et al., 2008; Gupta et al., 2008; Ioshikhes, Albert, Zanton, & Pugh, 2006; Segal et al., 2006; Yuan & Liu, 2008). The preferences for certain sequence motifs to occur at particular
locations within the nucleosome and the preferences for other motifs to be absent anywhere within a nucleosome are regarded as positioning signals (Segal & Widom, 2009). Nevertheless, some recent works dispute the leading role of DNA in determining nucleosome positioning in vivo (Peckham et al., 2007; Tolstorukov, Kharchenko, Goldman, Kingston, & Park, 2009; Valouev et al., 2008; Zhang et al., 2009). Without going into the details of the above-cited works and physical models (Miele, Vaillant, d’Aubenton-Carafa, Thermes, & Grange, 2008; Morozov et al., 2009; Tolstorukov, Colasanti, McCandlish, Olson, & Zhurkin, 2007), one can conclude that bioinformatics models can be used with high accuracy to predict general features of nucleosome organization (Tolkunov & Morozov, 2010).

Beyond the DNA sequence itself, many other cellular factors, such as DNA-binding proteins and chromatin remodeling complexes, influence nucleosome localization in the genome. Due to the impact of such factors, nucleosomes often do not adopt a single position. In a population of molecules, such positions can be quite variable or “fuzzy” (Zhang & Pugh, 2011). Because nucleosomes are nonoverlapping “beads-on-a-string,” the position of one nucleosome restricts the possible positions of adjacent nucleosomes by steric hindrance.

Tandem repeats arranged in long arrays with similar size and nucleotide structure are among the main classes of DNA sequence in eukaryotic genomes. Their proportion is higher in species with large genomes, particularly in plants. For example, tandem repeats account for approximately 15% of the large rye genome (Bedbrook, Jones, O’Del, Tompson, & Flavell, 1980). However, the nucleosome positioning in DNA tandem arrays has received insufficient attention, mostly because of the difficulties encountered in sequencing such genome regions. Several papers published before the genomic era on the characterization of nucleosome positioning show that tandem repeats display wide variation in the magnitude of intrinsic preferences in tandem DNA (Fitzgerald, Drydent, Bronson, Williams, & Anderson, 1994; Martinez-Balbas et al., 1990). Several factors make the genome regions that house long arrays of tandemly arranged sequences, particularly interesting for studying the effects of specific DNA structural features on nucleosome positioning. Similar to regularly spaced nucleosomes in chromatin, long arrays of tandem repeat DNA consist of regularly alternating monomers with almost identical nucleotide sequences. This property allows these arrays to be regarded as extended matrices of up to several mega base pairs. As was repeatedly noticed in the 1990s, two preferential monomer sizes of satellite DNAs in dicot plants are 165–190 and 330–380 bp; i.e., the monomer length (ML) tends to match the nucleosome repeat length (NRL) or be a multiple of it (Fitzgerald et al., 1994; Hemleben, Zentgraf, King, Borisjuk, & Schweizer, 1992; Kubis, Schmidt, & Heslop-Harrison, 1998). This trend was verified by an analysis of the database of tandem DNA repeats from plant species, PlantSat (Macas, Meszaros, & Nouzova, 2002). However, there are no suggestions regarding which factors could cause such a pattern or maintain it during evolution.

In this work, we have attempted to characterize the DNA sequences of tandem repeat families with different MLs, with the goal of clarifying the possible role of ML and nucleotide context in nucleosome formation potential. For this purpose, all tandem repeat families from PlantSat were divided into two classes: those with ML equal to NRL or a multiple of it (class 1) and those with ML not equal to or a multiple of NRL (class 2). We have developed the Phase method for predicting nucleosome formation potential, which applies an alphabet of all periodic (10–11 bp) and aperiodic dinucleotides as independent symbols. This method makes it possible to assess the effect of even small changes in the distribution of periodic dinucleotides in potential nucleosome formation sites (NFSs) relative to expectation. We have demonstrated that the ratio of ML to NRL correlates with the influence of nucleotide context on nucleosome positioning, which is reflected in a more regular type of nucleosome arrangement. Based on ML and nucleotide context, three main types of nucleosome arrangement in DNA tandem repeat arrays have been distinguished among a large diversity of nucleosome positioning profiles.

**Methods**

**A method for predicting NFSs**

We have developed the Phase method for recognizing NFSs. The training set of NFSs was taken from the Nucleosome Positioning Region Database (Levitsky, Katokhin, Podkolodnaya, Furman, & Kolchanov, 2005). The Phase method is described in detail in the Supplementary Data, Method 1. Briefly, the Phase method includes two steps: (a) calculation of the position weight matrix (PWM) profile (Levitsky et al., 2007) and (b) wavelet transformation (Lu, Liu, Xue, & Wang, 2004) of the PWM profile. The extended alphabet has been used for the PWM construction. This alphabet comprises all periodic and aperiodic dinucleotides as independent symbols. The dinucleotide pairs of the same type that are located at a distance of 10–11 nt (Supplementary Figure S1) are regarded as periodic (or phased). The weight value of any periodic or aperiodic dinucleotide in position \( i \) is determined by the probability of its occurrence in nucleosomal DNA in the training sample in position \( i \) and by the probability of its occurrence in random sequence with the same dinucleotide composition. The PWM prediction profile was used to identify peaks by the continuous wavelet transform (CWT) technique (Lu et al., 2004), since the peaks are expected to occur...
on every NRL. As a result, CWT indicates the profile regions (peaks) that reflect the pattern “linker DNA – nucleosome site – linker DNA” and represents the quantitative characteristics of the pattern.

**Determining the accuracy of prediction methods**

The accuracy of the Phase method has been compared with the accuracy of one of the most well-known analogous approaches (Segal et al., 2006) using control samples containing 1000 randomly selected human nucleosomal DNA sequences with lengths of 140–200 bp and linker DNA sequences with lengths of 100–200 bp (Dennis et al., 2007; Gupta et al., 2008; Tanaka & Nalaí, 2009, http://www.hgc.jp/~yanaka/assess2009/index.html). The procedure used for the comparison of these methods is described in detail in Supplementary Data, Method 2. The comparison has shown higher accuracy of the Phase method compared with the method of Segal et al. (2006) (Supplementary Figure S2).

**Calculating the threshold Phase score sufficient for nucleosome formation**

The comparison of Phase profiles for different nucleotide sequences is based on the analysis of the local maxima (heights of peaks) that correspond to individual NFSs. To obtain a reliable estimate for a threshold $\eta$ of peak height that is sufficient for context-dependent nucleosome formation, the Phase scores were computed for the complete rice and human genomes, and these genomes were downloaded from http://ftp.gramene.org/archives/ and http://hgdownload.cse.ucsc.edu/downloads.html#human, respectively. The median values of peak height for both genomes were approximately 0.013; accordingly, this value was chosen as the threshold height, $\eta$. The estimate for the fraction of context-dependent nucleosome positions in the whole genome is 50%, according to Kaplan et al. (2009). Based on this estimate, we suggested that at least 50% of our lowest peaks, corresponding to the weakest contextual nucleosome formation signals, lack sufficient contextual specificity. Therefore, the peaks with a height below the threshold value $\eta$ were regarded as insufficient for context-dependent nucleosome formation.

**Data acquisition and processing**

The nucleotide sequences of 975 monomers belonging to 165 families of plant tandem DNA repeats were extracted from the PlantSat database (Macas et al., 2002, w3lanc.umbc.cas.cz/PlantSat/). In addition, the sequences of 62 monomers belonging to the $pSc200$ family, which were isolated from various cereal species, and of three monomers from the $pSc250$ family from rye were extracted from different databases (Vershinin, Schwarczacher, & Heslop-Harrison, 1995). Before further analysis, the families were filtered to remove those with poor sequence quality or atypical monomers containing adjacent nontandem genomic sequences. There were four such cases. Of the remaining 161 families, 73 were represented by one monomer, 37 contained 1–5 monomers, and 51 contained more than five monomers.

Generally, we obtained Phase profiles for 3–4 kb tandem arrays of monomers, each containing several monomer copies. The local maxima can be found in such Phase profiles. Their values were estimated by a threshold of $\eta=0.013$. The highest peak in the profile is referred to as the dominant peak.

**Results**

**Deviations of the observed extrema of Phase scores from expected values**

The Phase method produced profiles of recognizable NFSs, with maxima corresponding to the locations of nucleosome core particles and minima corresponding to linker DNA. The first question that interested us was whether there are differences in the distributions of the maxima and minima (extrema) of observed and expected Phase profiles between the given DNA sequence from PlantSat (Macas et al., 2002) and a random sequence with the same dinucleotide content, in which any specific nucleotide context for nucleosome positioning should be absent. The average values of the maxima and minima of the Phase score profiles were calculated for all families, except those having an ML shorter than the minimal NRL size (155 nt) because they have absolute magnitudes for their extrema that are significantly lower than the threshold $\eta=0.013$ (see Methods). Selecting for minimal NRL size reduced the number of families analyzed to 124. For statistical comparison, the 25th, 50th, and 75th percentiles were computed for the distributions of the expected values of the maxima and minima. The results of the comparison between the observed and expected distributions of extrema are shown in Figure 1. Following the percentile definition, one-quarter of the analyzed families (31 out of 124) should have extrema values below the 25th percentile, and the same number of families should have extrema values above the 75th percentile. As seen in Figure 1, the maxima of deviations are detected for the 0–25th and 75–100th percentiles; the values have a positive trend there, whereas they have a negative trend in the interval between the 25th and 75th percentiles. For example, 51 families have maximum values above the 75th percentile. The differences between distributions are statistically significant according to a $\chi^2$ test ($p<0.002$ for minima, $p<1\times10^{-8}$ for maxima). Therefore, we have taken maximum value into account as one of the basic characteristic of a Phase profile in the classification of families according to nucleosome arrangement.
Heterogeneity of nucleosomal DNA profiles within individual tandem repeat families

Because tandem repeat families are, as a rule, multi-copy families, a question arises as to whether the Phase profiles produced by different copies of the same family are identical. To answer this question, the profiles for individual monomers were computed. To estimate the variation within a family, a sample of random sequences of the same length and dinucleotide content was constructed for each monomer. This sample was used to determine the expected maximum and minimum values of the profile. Respective expectations for a family were computed as the mean values for all random monomers.

Figure 2 shows examples of individual families that illustrate the differences between the profiles for individual monomer sequences. In some families, there can be relatively small difference between all the monomers (Figure 2(A)). In other cases, such as the family Nicotiana tomentosiformis NTRS, the profiles of some monomers are similar (Figure 2(B): U51848 vs. Z50790), and the profiles of other monomers are more different (Figure 2(B): Y08542 vs. Z50790). Families with 91–92% sequence identity are characterized by both groups of monomers. The heterogeneity of the monomers within families was taken into account when working out the criteria for the main types of nucleosome positioning.

Distribution of tandem repeat families according to ML

The 161 tandem DNA repeat families from the PlantSat database were divided into two classes based on the ratio of ML to NRL. This criterion was selected based on the assumption of NRL and ML coevolution (Hemleben et al., 1992) and on an analysis of the PlantSat database release (Macas et al., 2002). It was assumed that the NRL varied in the range of 155–205 bp. According to the chosen criterion, class 1 contained the families with MLs corresponding to an integer number of NRLs, and it included families with the following monomer sizes: 155–205, 310–410, 465–615, and 620–810 bp. Class 2 contained the families with MLs shorter than one NRL (ML < 155 bp) and families in which ML is greater than NRL but not an integer multiple of NRL, and it included families with the following monomer sizes: 1–154, 206–309, and 410–464 bp. Seven families of monomers longer than 820 bp were assigned to class 1. There were
almost twice as many families in class 1 as in class 2 (105 vs. 56).

**Main types of nucleosome arrangement in classes of tandem repeat families**

An important condition for the robustness of a quantitative estimation of nucleosome positioning in Phase profiles is the identification of peak shapes and the determination of peak number. We have detected diverse profiles, which indicate a multitude of possibilities for monomer sequences to form nucleosomes.

Analyzing the numbers and heights of the peaks in the monomers of all of the families from PlantSat, we identified certain general patterns. We propose a classification system for the main types of nucleosome arrangement in tandem arrays of monomers. Four features were taken into account in this classification, namely, (a) the ratio of ML to NRL; (b) the ratio of the number of peaks in the profile, \( N_{\text{PEAK}} \), to the number of nucleosome repeats in the monomer \( N_{\text{NR}} \); (c) the fraction \( d(f) \) of monomers in the family with the height of the dominant peak exceeding the threshold level \( \eta \); and (d) the height of the lowest peak in the profile (\( \eta_{\text{MIN}} \)), if the profile contains two or more peaks. The parameters ML, \( N_{\text{PEAK}} \), and \( \eta_{\text{MIN}} \) were estimated as average values for all monomers of a given family.

The following three types of nucleosome arrangement were defined using the criteria based on the above features. The **regular type** is determined by the following criteria: (a) ML is a multiple of NRL; (b) \( N_{\text{NR}} = N_{\text{PEAK}} \); (c) \( d(f) > 0.67 \); and (d) at ML \( \geq 2\text{NRL}, \eta_{\text{MIN}} > \eta/2 \). Two types were distinguished among the families that do not meet the above criteria, namely, a partially regular (or partial) type with \( d(f) > 0.33 \) and a flexible type with \( d(f) < 0.33 \).

As follows from condition (a), tandem repeats with regular type nucleosome arrangement can be assigned to class 1. Figure 3(A) shows an example of a regular profile for the family *Trifolium_TrR350*, with the ML matching two nucleosome repeats. The profile contains two pronounced peaks. The partial and flexible types are observed in monomers belonging to both classes. The partial type is similar to the regular type in that the profile contains at least one peak with a height above \( \eta \) (a dominant peak, Figure 3(B)). However, unlike a regular type, one of the four above criteria is not met in this case. An example illustrating noncompliance with criterion (b) is shown in Figure 3(B) for the family *pSc200* (class 1). Because there are several possibilities for noncompliance, the partial pattern is the most widespread type of nucleosome arrangement for monomers with ML \( \geq \text{NRL} \) (Figure 3(C)). In this case, criterion (a) for the regular type is not met.

Flexible type of nucleosome arrangement typically has multiple local maxima, indicating numerous potential nucleosome positions. An example of the flexible type is shown in Figure 3(D); this type is completely determined by the low height of the dominant peak in the profile \((0.009 < \eta)\), which is interpreted as a nucleotide context having insufficient potential to determine nucleosome formation.

Seven families containing five or more NRL are of special interest. According to the described classification, these families belong to regular and partial types. Actually, due to their exceptionally long MLs, the corresponding profiles are complex and nonuniform (mosaic). Figure 3(E) shows the family *Anemone_blanda_AbS1* with an ML of 1639 bp, which corresponds to eight to nine nucleosome repeats. The monomer profile falls into segments corresponding to the three main types of nucleosome arrangement. Consequently, we categorized the tandem repeat families having five or more NRL as having a mosaic type of arrangement that cannot be classified as one of the main types.

Thus, do the obtained results actually reflect the specific features of nucleosome positioning code rather than being the consequence of specific features of our method Phase? To clarify this issue, we repeated the above described computations using a method based on some structural features of DNA (Liu, Duan, Yu, & Sun, 2011), which we designated the Helix Curvature. See the Supplement, section “Comparison of the classifications obtained by different methods for NFS recognition” for detailed description of the computations and results of comparison. Comparison of the classifications obtained by applying the Phase and Helix Curvature methods has demonstrated that the numbers of families in classes 1 and 2 that belong to the same nucleosome arrangement type is considerably higher as compared to the number of such matches expected randomly. An incomplete overlapping of the corresponding Phase and Helix Curvature predictions, on the one hand, reflects the specific features of each method and, on the other, limits applicability of the methods. The Phase method predicts the nucleosome arrangement type based on dinucleotide periodicity, while the Helix Curvature based on the DNA helix curvature; however, all the DNA characteristics are inherently present at different ratios. We believe that our approach is best applicable to detecting the nucleosome arrangement type in the class of tandemly repeated DNA with the monomers of the length comparable to NRL (namely, varying in the range of 100–1000 bp) and organized into extended arrays.

**Quantitative estimates for different types of nucleosome arrangement**

Quantitative estimates for the different types of nucleosome arrangement observed in the 161 families and the two classes are shown in Figure 4. In addition, further data also reflect the relationship between the ML and type of nucleosome arrangement. Altogether, 55 families...
have ML equal to NRL. A large fraction of these families show the regular type of nucleosome arrangement, 26 out of 55, or 47%. The regular type is also observed for families with MLs amounting to two or more NRLs; however, the fraction of these families is considerably lower (14 out of 43 or 32.5%). Correspondingly, the fraction of class 1 families with partial type of nucleosome arrangement increases with ML (from 16% for families with ML equal to NRL to 51% for families with ML amounting to two or more NRLs). The partial type

Figure 3. Examples of Phase score profiles illustrating the main types of nucleosome arrangement in monomers. The X-axis shows the position relative to the start of the monomer and the Y-axis shows the Phase score. Fragments of the profiles for two monomers are shown. The dashed ovals denote the predictions of the most probable nucleosome positions (except for panel D). The dotted ovals in panels B and C show the predicted fuzzy nucleosome positions. EMBL accession numbers, MLs and classes are designated in the header. A – regular type, family *Trifolium TrR350*; B – partially regular type, family *pSc200*; C – partially regular type, family *Nicotiana tomentosiformis NTRS*; D – flexible type, family *Oryza TrsC*; and E – mosaic type, family *Anemone blanda AbS1*. The arrows mark the regions ascribed to the regular, partially regular, and flexible types of nucleosome arrangement.
is prevalent (84%) in families of class 2 that have an ML longer than one NRL. The overwhelming majority of class 2 families with ML < 155 (35 of 37) display a flexible type of nucleosome arrangement, which is most likely due to short ML because this proportion is significantly lower among the families with ML > NRL (3 out of 20 families). Nonetheless, two families with a partial type of nucleosome arrangement appear when the ML approaches to the minimal size of NRL (155 nt). Thus, the ML and type of nucleosome arrangement are related features; however, there is no universal trend in this association. Apparently, a minimal length (155 nt) of monomer is required for nucleosome formation potential.

**Deviations of average extrema of Phase scores in three types of nucleosome arrangement**

How are the revealed deviations in the distribution from the absolute magnitudes of the extrema (Figure 1) associated with the main types of nucleosome arrangement? Figure 5 partitions the general pattern of these deviations (Figure 5(A)) with respect to the three main types of nucleosome arrangement (Figures 5(B)–(D)). A quantitative assessment of these patterns obtained for all families from PlantSat, except those with ML < NRL, is presented on Table 1. The number of families analyzed is 124. Thirty out of 40 families of the regular type have maxima that exceed the expected values, but of the 30 families with flexible type arrangement and ML > NRL, only four families show a similar pattern. The 47 families of the partial type split into two approximately equal fractions: 23 families have a maximum exceeding the expected value, and 24 families have a maximum below the expected value. Thus, these data demonstrate that the pattern of extrema deviations is coupled to the main types of nucleosome arrangement. A tendency to exceed the expected magnitude of extrema is characteristic of the regular type, whereas the flexible type shows the opposite trend.

**Discussion**

**The choice of criterion for classifying tandem repeat families into two classes**

Plant genomes are especially enriched for tandem repeats. The PlantSat database, which was analyzed in this work, comprises 165 tandem repeat families, with MLs ranging from 27 bp (family PCSR, *Pinus densiflora*) to 3984 bp (family E3900, *Secale cereale*). As was repeatedly noticed in the 1990s, the ML in tandem repeats tends to match the NRL or be a multiple of it (Fitzgerald et al., 1994; Hemleben et al., 1992; Kubis et al., 1998). Therefore, it would be logical to classify the families in PlantSat into two classes according to NRL and to compare these classes according to the contributions of ML and nucleotide context to nucleosome formation potential.

The choice of the boundary parameter corresponding to the nucleosome repeat is of critical importance. The current literature lacks any unambiguous definition for this value. The nucleosome repeat contains the nucleosome core particle DNA and has a rigidly fixed length of 146±2 bp; however, the length of the linker DNA is a species- and tissue-specific characteristic (Kornberg, 1977; van Holde, 1989). Several authors believe that the minimal size of the linker DNA is 20 bp (Franz & de Jong, 2011; van Holde, 1989; Wong, Victor, & Mozziconacci, 2007). However, nuclease treatment of the chromatin housing several tandem families yields a regular nucleosome ladder with 157-bp periodicity (Sykorova, Fajkus, Ito, & Fukui, 2001; Vershinin & Heslop-Harrison, 1998). According to Widom (1992), the linker DNA lengths found in nature are preferentially quantized, differing from each other by multiples of ~10 bp, and the shortest NRL amounts to 159 bp. In our case, the crucial significance for selecting the ML range corresponding to the NRL was the ML distribution in the sample of families from PlantSat (Macas et al., 2002). A well-defined maximum of this distribution is observed in the range of 155–205 bp, which is in agreement with the aforementioned assumption that monomer size typically is a multiple of NRL. This observation encouraged us to select this range as a basis for class 1 families, in which the ML corresponds to an integer multiple of a single NRL. The families that do not meet this criterion were as assigned to class 2.

**Phase method distinguishes interfamily differences and intrafamily variability in the influence of monomer nucleotide context on nucleosome formation potential**

The periodic distribution of dinucleotides is thought to be related to sequence-dependent nucleosome formation (Bettecken & Trifonov, 2009; Cui & Zhurkin, 2010; Ioshikhes, et al., 1996; Ioshikhes, Hosid, & Pugh, 2011; Kaplan et al., 2009). The specific features of this distribution along the monomers of the tandem repeat families...
can be elucidated by analysis of nucleosome positioning profiles using the Phase method. The Phase method, like other methods (Kaplan et al., 2009; Segal et al., 2006), applies a PWM framework. However, we do not consider the distribution of dinucleotide frequencies. Instead,

we calculated the PWM score by taking into account the distributions of periodic and aperiodic dinucleotides. We use this approach because the PlantSat database represents a wide range of plant species, and it is likely that all combinations of periodic (phased) dinucleotides contribute in some way, in a species-specific manner, to the nucleosome positioning (Bettecken & Trifonov, 2009; Gabdank, Barash, & Trifonov, 2009; Salih, Salih, & Trifonov, 2008; Travers, Hiriart, Churcher, Caserta, & Di Mauro, 2010). We compared the dinucleotide distributions to the expected values of the maxima and minima of random sequences that have the same lengths and dinucleotide compositions.

Figure 5. Patterns of scatterplot deviations between the observed and expected average values of maxima and minima of Phase profiles in the main types of nucleosome arrangement. A – pattern obtained for all families from PlantSat, except those with ML < NRL. B – pattern of families that have a regular-type arrangement. C – partially regular type. D – flexible type. ○ – observed values for families; ▲ – expected values for the random sequence sets that have the same lengths and dinucleotide compositions.

Table 1. Number of families with particular combinations of average maximum Phase score deviation and nucleosome arrangement type.

|               | Regular type | Partial type | Flexible type |
|---------------|--------------|--------------|--------------|
| MaxOBS > MaxEXP | 30           | 23           | 4            |
| MaxOBS < MaxEXP | 10           | 24           | 26           |

Notes: $M_{OBS}$ – an observed average maximum Phase score for given family. $M_{EXP}$ – an expected average maximum Phase score for a random family that has the same length and dinucleotide composition. According to Yates corrected $\chi^2$ test for $2 \times 2$ contingency tables, the relationships of data for the following pairs are significant: Regular type–Partial type, Regular type–Flexible type, Partial type–Flexible type ($p<0.05$, $p<2E-6$ and $p<4E-3$, respectively).

122

V.G. Levitsky et al.
We believe that this approach is the most optimal tandem repeat analysis method that has been confirmed by comparison its recognition accuracy (Supplementary, Section 2, Table S1, Figure S2) with one of the most widely used methods (Segal et al., 2006).

Randomly shuffled sequences are routinely used in sequence analysis to evaluate the statistical significance of a biological sequence (Jiang, Anderson, Gillespie, & Mayne, 2008). Because the Phase method is based on dinucleotide statistics, we used a first-order Markov chain to generate sets of random sequences with the same dinucleotide composition. Based on patterns in the distribution of observed extrema with respect to expected extrema (Figures 1 and 5), we assume that the observed extrema that are higher than expected indicate a nucleotide context that favors nucleosome formation and that the observed extrema that are lower than expected indicate weak influence of nucleotide context. Apparently, the revealed deviations reflect the influence of different evolutionary factors pertaining to the formation of nucleosomes in different nucleotide contexts and in the presence of different tandem repeat families. The great variety of patterns observed suggests that the tandem repeat families are very distinct in terms of their extent of exposure to the evolutionary factors. Comparison of the nucleosome positioning profiles of individual monomers within the same family by the Phase method (Figure 2) demonstrated the heterogeneity of the monomers. It has been shown that natural selection may act independently on different DNA sequence properties responsible for local chromatin organization (Babbitt, Tolstorukov, & Kim, 2010). Our results confirmed the idea that there are different evolutionary rates for nucleotide context characteristics in different tandem repeat families.

**Main types of nucleosome arrangement along tandem repeat arrays**

In addition to the ratio of ML to NRL, the peak number and peak height of the Phase profiles and the intrafamily heterogeneity of the monomer populations were taken into account when elaborating the criteria for the classification of nucleosome positioning patterns (Figure 3). The scheme presented in Figure 6 illustrates the positional relationships of the adjacent nucleosomes resulting from the proposed main types of nucleosome arrangement. In the case of the regular type, the positions of all of the nucleosomes along an array of monomers become uniquely determined (Figures 3(A), 6(A)), and all of the copies of the monomers of a given family should be congruent with their nucleosome positioning. The regular type is unfeasible in class 2 families, as the presence of a nucleosome at a given position within one monomer interferes with its presence at the same position in adjacent monomers due to the mismatch of ML and NRL (Figures 3(C), 6(C)).

Several class 1 families, such as pSc200 of *S. cereale*, display only one prominent peak in their profiles and, correspondingly, have partially regular nucleosome arrangement (Figure 3(B)). Such a profile suggests that the position of only one key nucleosome (Kiyama & Trifonov, 2002; Liu & Stein, 1997) out of several in a monomer is determined in strict accordance with the specific nucleotide context features (Figures 6(B), (C)). The positions of other nucleosomes can be determined based on probabilistic models, similar to the statistical positioning of nucleosomes (Mavrich et al., 2008), which implies that the nucleotide sequence allows fuzzy nucleosome positioning between two neighboring key nucleosomes. According to the hypothesis of mosaic nucleosome arrangement in chromatin (Liu & Stein, 1997), one initially formed nucleosome can control the formation of an array of up to 10 nucleosomes in the direct vicinity.

The freedom of nucleosome positioning within monomers is also determined by the number of nucleosome repeats that can be accommodated between the adjacent key nucleosomes. We assume that the differences between the regular and partially regular types are not too large in cases that are similar to pSc200 (Figures 6(A), (B)). Experiments involving micrococcal nuclease treatment of chromatin and subsequent hybridization to a pSc200 probe gave a regular pattern of oligomeric hybridization fragments with lengths that were multiples of the nucleosome repeat, yet with somewhat fuzzy boundaries (Vershinin & Heslop-Harrison, 1998). The partial type arrangement presented in Figure 6(B) is an inherent feature of class 1 families and is characterized by a limited number of fuzzy nucleosomes. Every monomer along the tandem array contains a key nucleosome.

The partially regular positioning profile that is typical for class 2 families has some specific features (Figure 6 (C)). Here, as a rule, the monomers contain several overlapping nucleosome positioning sites that can determine different phases of nucleosome positioning. For example, the presence of a set of preferred positioning signals within the monomer, which can be selected by key nucleosomes, has been experimentally shown for the NTRS family (Matyasek, Gazdova, Fajkus, & Bezdek, 1997, Figure 3(C)). Owing to a mismatch between the ML and NRL and the heterogeneity of the monomer sequences, there is not a key nucleosome in every monomer, and linker DNA can be of variable length (Figure 6 (C)). Thus, a variable number of fuzzy nucleosomes may be placed between adjacent key nucleosomes.

The absence of a pronounced peak in the profile provides maximum freedom in choosing the sites for nucleosome positioning. In this case, probabilistic models are applicable to the whole array of monomers (Figures 3(D) and 6(D)). This flexible type of nucleosome arrangement is particularly characteristic of the families belonging to class 2 that have short MLs.
The fraction of families that have a flexible type arrangement decreases as ML increases. This observation suggests that the evolutionary establishment of a nucleotide context with a high preference for key nucleosomes is less likely in the DNA sequence of short monomers compared with longer monomers.

One of the fundamental differences between the regular positioning type and both the partial and flexible types is that the lengths of the linker DNA and, correspondingly, the NRL, in the former case remain constant within the entire array of monomers, whereas these parameters may vary even between the adjacent monomers in the latter types (Figure 6). Nucleosomes dynamically interconvert into more compact units, including a fiber that is 30 nm in diameter, which further folds into higher order structures. According to the models of chromatin internal structure, different NRLs influence the orientation of the linker histones and correspond to different fiber conformations (Robinson, Fairall, Huynh, & Rhodes, 2006; Wong et al., 2007). Presumably, the main types of nucleosome arrangement are associated with the formation of different types of chromatin structural organization. For example, it has been found that nucleosome repeats with a constant length of approximately 167 bp are involved in the formation of a classic solenoid-type chromatin conformation (Wong et al., 2007). In our case, a majority of families that have similar ML size form a regular type of nucleosome arrangement, which accounts for a quarter (24%) of all the families compiled in PlantSat (Figure 4).

Our proposed classification reflects a process of nucleotide context accumulation during evolution, wherein different tandem arrays are formed from monomers that have different lengths. Apparently, the tandem repeat family formed a regular type of nucleosome arrangement, which favored nucleosome formation and led to selection for an ML that is equal to or a multiple of NRL. If the strengths and mode of selection associated with nucleosome positioning are weak or negative for a given family, then the favored arrangements are the partial and flexible types, respectively. These trends are illustrated by the different patterns of deviation distributions for the observed maxima and minima and the expected maxima and minima in the main types of nucleosome arrangement (Figure 5). Deletion of the nucleotide context that we produce in random sequences resulted in a reduction of the Phase score for the regular-type families and an increase in Phase score for flexible-type families. These data confirm the role of nucleotide context in nucleosome arrangement. Recently, evidence for both positive and negative selection linked to human nucleosome positioning was obtained by Prendergast and

Figure 6. Schematic representation of the main types of nucleosome arrangement in tandem arrays of monomers with different ratios of ML to NRL. (A) Regular type, class 1, ML/NRL = 1; (B) Partial type, class 1, ML = 2NRL; (C) Partial type, class 2, NRL < ML < 2NRL, variable linker length; and (D) the nucleosome positions are chosen arbitrarily, as the signals for preferential nucleosome positioning are absent. In panels A, B, and C it is assumed that the monomer-specific context-directed nucleosome localization signal in each of these three variants is equal in its strength. The “red smiley”/“blue cross” nucleosome positions mark signals that are sufficient/not sufficient for nucleosome formation. These cases are predicted by Phase scores that are greater/less than the threshold $\eta = 0.013$. 

(ML < NRL).
Semple (2011). According to that study, selection appears to be acting on particular base substitutions in the nucleosome core and linker regions.

Acknowledgments

The authors thank Drs. V.J. Makeev, I.B. Rogozin, and S.A. Demakov, for critical review of the manuscript. This work was supported by the Russian Foundation for Basic Research (grant 11-04-01888 to V.G.L. and grant 12-04-00512 to A.V.V.), Program of the Russian Academy of Sciences 6.8 “Molecular and Cell Biology” (to V.G.L.), and the Siberian Branch of the Russian Academy of Sciences (Integration project 51 to A.V.V.).

References

Babbitt, G. A., Tolstorkov, M. Y., & Kim, Y. (2010). The molecular evolution of nucleosome positioning through sequence-dependent deformation of the DNA polymer. Journal of Biomolecular Structure & Dynamics, 27, 765–780.

Bedbrook, J. R., Jones, J., O’Del, M., Tompson, R., & Flavell, R. (1980). A molecular description of telomeric heterochromatin in Secale species. Cell, 19, 545–560.

Bettecken, T., & Trifonov, E. N. (2009). Repertoires of the nucleosome-positioning dinucleotides. PLoS ONE, 4, e7654.

Cui, F., & Zhurkin, V. B. (2010). Structure-based analysis of DNA sequence patterns guiding nucleosome positioning in vitro. Journal of Biomolecular Structure & Dynamics, 27, 821–841.

Dennis, J. H., Fan, H., Reynolds, S. M., Yuan, G., Meldrim, J. C., Richter, D. J., … Kingston, R. E. (2007). Independent and complementary methods for large-scale structural analysis of mammalian chromatin. Genome Research, 17, 928–939.

Drew, H. R., & Travers, A. A. (1985). DNA bending and its relation to nucleosome positioning. Journal of Molecular Biology, 186, 773–790.

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I. K., Sharon, E., Lubling, Y., … Segal, E. (2008). Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Computational Biology, 4, e1000216.

Fitzgerald, D. J., Drydent, G. L., Bronson, E. C., Williams, J. S., & Anderson, J. N. (1994). Conserved patterns of bending in satellite and nucleosome positioning DNA. Journal of Biological Chemistry, 269, 21303–21314.

Franz, P., & de Jong, H. (2011). From nucleosome to chromosome: A dynamic organization of genetic information. The Plant Journal, 66, 4–17.

Gabdank, I., Barash, D., & Trifonov, E. N. (2009). Nucleosome DNA bendability matrix (C. elegans). Journal of Biomolecular Structure & Dynamics, 26, 403–411.

Gupta, S., Dennis, J., Thurman, R. E., Kingston, R., Stamato-yannopoulos, J. A., & Noble, W. S. (2008). Predicting human nucleosome occupancy from primary sequence. PLoS Computational Biology, 4, e1000134.

Hernießen, V., Zentgraf, U., King, K., Borisjuk, N., & Schweizer, G. (1992). Middle repetitive and highly repetitive sequences detect polymorphisms in plants. In G. Kahl, H. Appelhans, J. Kompf, & A. J. Driesel (Eds.), DNA-polymorphisms in eukaryotic genomes, Vol 5 (pp. 157–170). Heidelberg: BioTechForum (BTF), Advances in Molecular Genetics, Huethig.

Ioshikhes, I., Albert, I., Zanton, S. J., & Pugh, B. F. (2006). Nucleosome positions predicted through comparative genomics. Nature Genetics, 38, 1210–1215.

Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M., & Trifonov, E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. Journal of Molecular Biology, 262, 129–139.

Ioshikhes, I., Hosid, S., & Pugh, B. F. (2011). Variety of genomic DNA patterns for nucleosome positioning. Genome Research, 21, 1863–1871.

Jiang, M., Anderson, J., Gillespie, J., & Mayne, M. (2008). UShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinformatics, 9, 192.

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., … Segal, E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature, 458, 362–366.

Kiyama, R., & Trifonov, E. N. (2002). What positions nucleosomes?–A model FEBS Let., 523, 7–11.

Kornberg, R. D. (1977). Structure of chromatin. Annual Review of Biochemistry, 46, 931–954.

Kubis, S., Schmidt, T., & Heslop-Harrison, J. S. (1998). Repetitive DNA elements as a major component of plant genomes. Annals of Botany, 82(Suppl 1), 45–58.

Levitsky, V. G., Ignatieva, E. V., Ananko, E. A., Turnaev, I. I., Merkulova, T. I., Kolchanov, N. A., & Hodgman, T. C. (2007). Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. BMC Bioinformatics, 8, 481.

Levitsky, V. G., Katokhin, A. V., Podkolodnaya, O. A., Furrman, D. P., & Kolchanov, N. A. (2005). NPRD: Nucleosome positioning region database. Nucleic Acids Research, 33, 67–70.

Liu, H., Duan, X., Yu, S., & Sun, X. (2011). Analysis of nucleosome positioning determined by DNA helix curvature in the human genome. BMC Genomics, 12, 72.

Liu, K., & Stein, A. (1997). DNA sequence encodes information for nucleosome array formation. Journal of Molecular Biology, 270, 559–573.

Lu, X. Q., Liu, H. D., Xue, Z. H., & Wang, X. Q. (2004). Maximum spectrum of continuous wavelet transform and its application in resolving an overlapped signal. Journal of Chemical Information and Computer Sciences, 44, 1228–1237.

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature, 389, 251–260.

Macas, J., Meszaros, T., & Nouzova, M. (2002). PlantSat: A specialized database for plant satellite repeats. Bioinformatics, 18, 28–35.

Martinez-Balbas, A., Rodriguez-Campos, A., Garcia-Ramirez, M., Sainz, J., Carrera, P., Aymami, J., & Azorin, F. (1990). Satellite DNAs contain sequences that induced curvature. Biochemistry, 29, 2342–2348.
