BIOINFORMATICS ORIGINAL PAPER
Vol. 30 no. 6 2014, pages 761–767
doi:10.1093/bioinformatics/btt621

Genome analysis
Advance Access publication October 31, 2013

NCS: incorporating positioning data to quantify nucleosome stability in yeast

Jung-Hsien Chiang* and Chan-Hsien Lin
Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 70101, Taiwan

Associate Editor: John Hancock

ABSTRACT

Motivation: With the spreading technique of mass sequencing, nucleosome positions and scores for their intensity have become available through several previous studies in yeast, but relatively few studies have specifically aimed to determine the score of nucleosome stability. Based on mass sequencing data, we proposed a nucleosome center score (NCS) for quantifying nucleosome stability by measuring shifts of the nucleosome center, and then mapping NCS scores to nucleosome positions in Brogaard *et al.*'s study.

Results: We demonstrated the efficiency of NCS by known preference of A/T-based tracts for nucleosome formation, and showed that central nucleosomal DNA is more sensitive to A/T-based tracts than outer regions, which corresponds to the central histone tetramer-dominated region. We also found significant flanking preference around nucleosomal DNA for A/T-based dinucleotides, suggesting that neighboring sequences could affect nucleosome stability. Finally, the difference between results of NCS and Brogaard *et al.*'s scores was addressed and discussed.

Contacts: jchiang@mail.ncku.edu.tw

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on April 9, 2013; revised on October 21, 2013; accepted on October 24, 2013

1 INTRODUCTION

The nucleosome is a fundamental repeating unit of DNA packaging in eukaryotic chromatin. In yeast, it is composed of a core containing ~147-bp segments of DNA wrapped around a histone octamer, which consists of two H2A–H2B dimers and one $(H3–H4)_2$ tetramer. Most of the DNA is wrapped in nucleosomes to resemble a string with beads, and neighboring nucleosomes are generally separated by linker DNA. Approximately 75–90% of genomic DNA is packaged in nucleosomes. A previous study indicated that in *Saccharomyces cerevisiae*, 81% of the genome sequences appear to be wrapped in nucleosomes (Lee *et al.*, 2007). The positioning of nucleosomes influences the accessibility of proteins to DNA or blocks transcription factor-binding sites. Consequently, nucleosomes are strongly associated with chromatin and gene expression, and they can regulate transcription initiation, DNA replication and DNA repair (Liu *et al.*, 2006; Li *et al.*, 2007). Nucleosome positions, combined with histone variants and histone modifications, are the key to genome regulation and gene expression.

In *S.cerevisiae*, an increasing number of studies have focused on nucleosome positioning by using several technologies over the past several years. Early studies for identifying nucleosome positions primarily relied on the tiled microarray hybridization approach. The first set of high-throughput positioning data was generated in 2005 (Yuan *et al.*, 2005), followed by the complete high-resolution map of nucleosome occupancy on a whole genome (Lee *et al.*, 2007). With the maturity of high-throughput DNA sequencing technology, many high-resolution maps of genome-wide nucleosome positions have been made in recent years (Albert *et al.*, 2007; Brogaard *et al.*, 2012; Field *et al.*, 2008; Jiang and Pugh, 2009; Kaplan *et al.*, 2009; Weiner *et al.*, 2010 Mavrich *et al.*, 2008; Rizzo *et al.*, 2011; Tsankov *et al.*, 2010; Tsui *et al.*, 2011; Xi *et al.*, 2011; Zhang *et al.*, 2009; Whitehouse *et al.*, 2007). Among those data, some nucleosomes have very consistent positions within the genome, but others are unstable and varied. Divergence between the data can be caused by both artificial and natural factors. The artificial factors, including the biological procedure of nucleosome digestion (Chung *et al.*, 2010), quality of high-throughput technology and computational method of data processing, make the comparisons between different datasets more complicated. As for natural factors, several determinants, such as distinct DNA sequences, DNA methylation, histone variants, post-translational modifications, transcription factors and chromatin remodelers, can affect nucleosome positions (Segal and Widom, 2009b). This phenomenon of divergence in nucleosome positions led to studies on the fuzziness or stability of nucleosome positions, such as 'position cluster' (Cole *et al.*, 2012). During processing of nucleosome positioning data, the nucleosome position and the intensity for that position can be calculated using several computational methods; the standard deviation of intensity values was usually directly used as fuzziness scores to measure the nucleosome dynamics or stability in some relevant studies (Albert *et al.*, 2007; Mavrich *et al.*, 2008). However, based on the design of biological procedures, such as different concentrations of micrococcal nuclease digestion, this standard deviation might not truly represent the dynamics or stability of nucleosome position, and could possibly result from digestion preference or bias, or could be affected by the binding strength between the histone and DNA, instead of nucleosome dynamics or stability.

In this study, we quantified nucleosome stability by considering the shift of nucleosome centers, treating nucleosome stability as the probability of a nucleosome center in a specific position, and determined the probability of shifting by observing differences between diverging data. Because high-throughput sequencing data have been easily available in recent years, we developed

*To whom correspondence should be addressed.

a computational approach using the Gaussian function to define the nucleosome center score (NCS) based on several published sequencing data of *S.cerevisiae* under different conditions. Then, we used both NCS and precise nucleosome positions at base-pair resolution (Brogaard *et al.*, 2012) to demonstrate the efficiency of NCS and to represent a clear relationship between NCS and sequence characteristics. With the help of high-resolution Brogaard *et al.*'s nucleosome data and a reliable NCS, clear patterns of dinucleotide preference around nucleosomal DNA and its neighboring DNA were illustrated in our study. Finally, we also indicated differences between NCS and scores from Brogaard *et al.*'s study, and discussed their significance.

2 METHODS

2.1 Datasets

In this study, we integrated several published nucleosome positioning data in *S.cerevisiae* to calculate NCS, and applied NCS on precise nucleosome positions. We then discussed NCS by sequence characteristics of nucleosomal DNA. All experimental data were collected from the following sources:

(1) Raw data of genome-wide sequencing for nucleosome positions were collected from the Sequence Read Archive database (Kodama *et al.*, 2012). We chose five studies (Kaplan *et al.*, 2009; Rizzo *et al.*, 2011; Tsui *et al.*, 2011; Weiner *et al.*, 2010; Xi *et al.*, 2011) with the same sequencing platform from Illumina Genome Analyzer I/II, but on different biological cultures, environments and micrococcal nuclease digestions (summarized in Table 1). In total, 30 sequencing datasets on nucleosome positions in *S.cerevisiae* were used in this study.

(2) Nucleosome positions at base-pair resolution were obtained from a previous article (Brogaard *et al.*, 2012). Brogaard *et al.*'s data are the first to locate nucleosome positions genome wide, with unprecedented detail and accuracy and were regarded as the precise well-defined nucleosome positions in our study.

Table 1. Summary of sequencing data for nucleosome positioning

Authors	Condition	SRA accession number[a]
Kaplan *et al.*	Wild-type	023800, 023801, 023802, 023803, 023804, 023805
	In galactose medium	023806, 023807, 023808
	In ethanol medium	023809, 023810, 023811, 023812
Weiner *et al.*	Wild-type	032450, 032451
	Heat shock	032452, 032453, 032454
Tsui *et al.*	Wild-type	063958
Xi *et al.*	Wild-type, partial digestion	191761, 191762
	Wild-type, complete digestion	090251, 090253
	Heat shock	090254, 090255, 090256, 090257
	Wild-type, crosslink	090258, 090259
Rizzo *et al.*	Wild-type	353537

Note: [a]The prefix of SRA accession number 'SRR' is ignored here.

(3) Reference genome sequences of *S.cerevisiae* were downloaded from the University of California, Santa Cruz Genome Bioinformatics site (Dreszer *et al.*, 2012). There are three versions for *S.cerevisiae*: sacCer1, sacCer2 and sacCer3, which were assembled in 2003, 2008 and 2011, respectively. We adopted the sacCer2 version, which was also used in Brogaard *et al.*'s study, for analysis.

(4) The information of open reading frame was obtained from the study of Xu *et al.* (2009), comprising 5171 verified or uncharacterized transcription start sites (TSSs).

2.2 Detection of nucleosome positions from sequencing data

For the 30 sequencing data downloaded from the Sequence Read Archive database, we used Burrows–Wheeler Aligner (BWA) (Li and Durbin, 2009), which is an efficient program that can align short read sequences against long reference genome sequences to map the nucleosomal DNA fragments against a reference genome of yeast. SAMtools (Li *et al.*, 2009) was then used to manipulate the BWA alignment format and generate alignments in a per-position format, such as the Browser Extensible Data (BED) format. After the BED format was prepared to record the location of short reads on the genome, nucleosome positions were estimated by Nucleosome Positioning from Sequencing (NPS) (Zhang *et al.*, 2008). NPS is a python software package that can identify peaks of qualifying width as positioned nucleosomes (PNs). We made the parameters in NPS more flexible to include as many nucleosomes as possible in each datum because further information on NCS was based on these nucleosome positions. According to the amount and quality of sequencing, most sequencing data can generate ~50 000 PNs in the genome. To ensure the essential quality of these 30 PN data, comparisons among PNs were made in the Supplementary Information.

2.3 Calculation of NCS of each base pair

Assuming the nucleosome can possibly be shifted in neighboring positions, and shifting occurs more easily when the shift distance is smaller, the data of nucleosome positioning can be considered as the balanced result of shifting. For each PN *i*, we applied the Gaussian function as the kernel probability density function to describe the nucleosome existence as follows:

$$p_i(x) = \sum_{j=1}^{m_i} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-c_j)^2}{2\sigma^2}\right) \tag{1}$$

where $p_i(x)$ is the probability of nucleosome center existence for nucleotide location x in the genome, and m_i is the total number of PNs in PN *i*; for each PN, c_j is the nucleosome center, and σ is the standard deviation of reasonable shift distances. According to the observed distribution of nucleosome shift distance on six duplicated data from Kaplan study *et al.*'s (2009), σ was set to 10 in our work.

Based on the 30 different PNs with different biological conditions, a linear combination of probabilities for nucleosome center existence on each datum was used to calculate the NCS for each base pair. For *n* PNs, the values of NCS for each base pair *x*, $NCS(x)$, was defined as:

$$NCS(x) = \sum_{i=1}^{n} w_i p_i(x) \tag{2}$$

where w_i is the weight for each data source or the confidence score for the data. Because our focus is mainly on the shift of nucleosome center rather than the intensity of the nucleosome, we assumed here that the quality of each nucleosome positioning datum is the same, and thus treated the 30 sets of nucleosome positioning data equally. The *R* implementing code for NCS is released in the Supplementary Data, and the original values of NCS for each base pair on chromosomes are shown in Supplementary Table S1.

2.4 Application of NCS for data of nucleosomes

NCS is the integration result of 30 PNs, and the value of the NCS is only a representation of neighboring nucleosome centers. For each base pair, a high NCS means that the nucleosome center is more stable, whereas a low NCS suggests an unstable nucleosome center that can easily be affected by other factors. Because we have a score scheme for nucleosome stability but no precise nucleosome positions data, nucleosome positions at base-pair resolution (Brogaard *et al.*, 2012) were used in this work. Although there have been several nucleosome positioning datasets on a whole genome in yeast since Lee *et al.*'s work (2007), no data could reach base-pair resolution until Brogaard *et al.*'s study. The significant periodical pattern of dinucleotide frequency in Brogaard *et al.*'s study indicates their precise nucleosome positions because even small noises can lead to a blurred dinucleotide pattern (Supplementary Fig. S4). Therefore, we applied NCS on Brogaard *et al.*'s data by mapping NCS values to the positions of nucleosome centers and used these values to represent nucleosome stability. Brogaard *et al.*'s data, with normalized NCS, are presented in Supplementary Table S2.

To show that incorporating multiple data (NCS) has an advantage over single datum in quantifying nucleosome stability, we took one PN as a reference for the analysis. The data with accession number SRR023800 (Kaplan *et al.*, 2009) were chosen to represent PNs. Because SRR023800 has occupancy scores only for each nucleosome but not for each base pair, the score cannot be directly applied to Brogaard *et al.*'s data the same as NCS. The scores of SRR023800 were applied to Brogaard *et al.*'s data by calculating the distance of any two nucleosome positions separately from SRR023800 and Brogaard *et al.*'s data, assuming two positions with a small distance as the same one and keeping these positions, then applying the SRR023800 score (~67.8% of nucleosomes in Brogaard *et al.*'s data were kept and applied by SRR023800 scores).

2.5 Evaluation of the relationship of nucleosome score with dinucleotide frequency pattern

To initially observe the difference among nucleosomes with different scores, the nucleosome data were classified into five classes in ascending order of score values, and the aggregate dinucleotide frequency was drawn separately on each class for viewing. To further represent the relationship between nucleosome score and dinucleotide frequency pattern, or to measure how pattern changes when score changes, we used the frequency change rate to describe the degree of dinucleotide frequency change when the nucleosome score increased. The frequency change rate for each dinucleotide type was measured as follows. First, the nucleosome data were classified into 20 classes based on their scores, and the aggregate dinucleotide frequency pattern was calculated for each class. Then, for each location around the nucleosome center, we used a linear regression to fit the change in dinucleotide frequencies among the 20 classes, and used the value of the slope from the linear fit to measure the dinucleotide frequency change rate. The pattern of frequency change rate indicates the preference or influence of different dinucleotide types on nucleosomal DNA.

To provide the degree of confidence on the linearity of frequency change rate, or to quantify the sensitivity of the dinucleotide feature to nucleosome score for each location around the nucleosome center, the sensitivity to dinucleotide was proposed. Combining all the coefficients of determination, *r*, from linear regression results for all dinucleotide types, the sensitivity to dinucleotide frequency change was defined as:

$$Sensitivity = \frac{\sum_{i=1}^{n} r_i |s_i|}{\sum_{i=1}^{n} |s_i|} \tag{3}$$

where *i* indicates the dinucleotide type, and s_i and r_i are the slope value and coefficient of determination (*r*-squared value) in linear regression, respectively. The values of sensitivity to dinucleotide were illustrated with the pattern of dinucleotide frequency rate.

3 RESULTS AND DISCUSSION

3.1 NCS shows clear discrimination of AA/TT/AT/TA dinucleotide frequencies

Previous studies suggest that nucleosome positions are highly related to sequence patterns *in vitro* and *in vivo* (Segal *et al.*, 2006; Kaplan *et al.*, 2009, 2010), and sequence information is a strong determinant for nucleosome formation and inhibition. Therefore, we used the position-dependent frequencies of AA/TT/AT/TA dinucleotides to illustrate the difference for nucleosome sets with different levels of nucleosome score, and presented the efficiency of NCS. As shown in Figure 1A, Brogaard *et al.*'s data were classified into five groups according to their NCP/noise scores defined in the original article (Brogaard *et al.*, 2012). The nucleosomes with a large NCP/noise score are well positioned, and stronger periodic dinucleotide sequence features are favored in nucleosomes having large NCP/noise, which is consistent with the original study. However, when classifying Brogaard *et al.*'s data according to NCS (depicted in Fig. 1B), we observed better discrimination of frequency patterns among each class. We also noticed a very distinct tendency for nucleosome sets with larger NCS to have lower relative AA/TT/AT/TA frequencies than those with smaller NCS, indicating that nucleosome positioning disfavors the A/T-based dinucleotides. This tendency is also supported by the result of Brogaard *et al.*'s data classification based on SRR023800 score (Fig. 1C) or SRR023800 classification (Supplementary Fig. S5).

The frequency patterns in Figure 1B and C imply that discrimination in the central region is clearer than the outer region, so the patterns of absolute AA/TT/AT/TA frequency change rate for NCS and SRR023800 were illustrated in Figure 2. The smoothing result for NCS depicts a symmetric pattern surrounding the nucleosome center inside −40 to +40 bp, and a decrease outside of the region −40 to +40 bp (drop from 0.429 to 0.009 in −40 to −70 bp and from 0.402 to 0.014 in +40 to +70 bp). Several previous studies (Thastrom *et al.*, 2004; Bohm *et al.*, 2011; van der Heijden *et al.*, 2012) indicate that nucleosome positioning is first accessed by core histone $(H3–H4)_2$ tetramers, which occupy ~70–80 bp surrounding the center, and the $(H3–H4)_2$ tetramers dominate the free energy of histone–DNA interaction. The high change rate inside the region of −40 to +40 bp, low change rate outside of the region and symmetric pattern for the central region of nucleosomal DNA (Fig. 2) strongly suggest that the mechanism of nucleosome stability and dynamics is mainly associated with the $(H3–H4)_2$ tetramer, rather than the whole nucleosome, and the dinucleotide of AA/TT/AT/TA could have more influence on the $(H3–H4)_2$ tetramer than H2A or H2B.

Because NCS shows its better discriminatory ability on dinucleotide frequency patterns for nucleosomes with different levels of scores in Figure 2 (*P*-value < 2.2e-16), and the difference in the central nucleosome region and outer region corresponds to the previous studies, NCS is supported by known biological phenomena. Moreover, based on the evidence that almost all PNs are consistent with results of NCS (data not shown), we believe

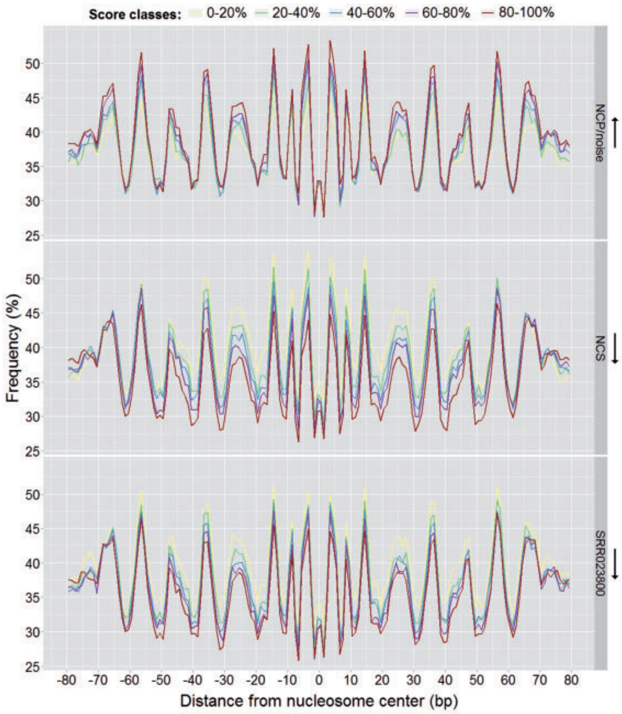

Fig. 1. Position-dependent frequencies of AA/TT/AT/TA dinucleotides for Brogaard's data were classified in ascending order of score values, based on (**A**) Brogaard's nucleosome score (NCP/noise), (**B**) NCS and (**C**) SRR023800 score. The right black arrow indicates the preference direction of frequencies versus score values. NCS has good discriminatory ability on dinucleotide frequency for each class, and shows disfavored behaviors on AA/TT/AT/TA dinucleotides, which is contrary to the result of Brogaard's score (NCP/noise)

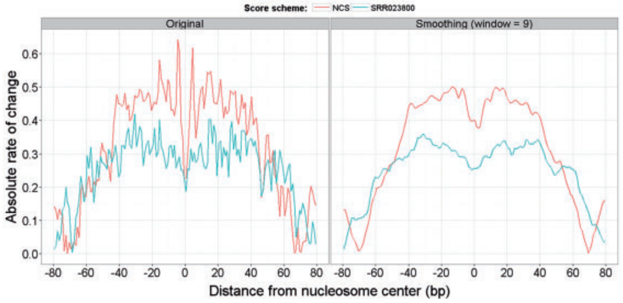

Fig. 2. Absolute change rate of AA/TT/AT/TA dinucleotide frequencies around the nucleosome centers for NCS and SRR023800 on Brogaard's data. The left figure shows the original absolute change rate, and the right figure is the smoothing result of moving average, with sliding window = 9 bp. The pattern is clearly depicted as having a symmetric high rate of change surrounding the center of the nucleosome, with a decrease on both sides (drop is 0.42 in −40 to −70 bp and 0.39 in +40 to +70 bp)

that results based on NCS are more accurate than NCP/noise in regard to nucleosome preference for AA/TT/AT/TA, and more details on the opposite condition will be discussed in later sections.

3.2 NCS reveals the flanking pattern on the preference of AA/TT, AT and TA dinucleotides

We showed using NCS can successfully discriminate the patterns of dinucleotide frequency for different nucleosome stabilities, and provided the general result of the A/T-based dinucleotides on nucleosomal DNA, to further investigate the preference of all types of dinucleotides simultaneously on nucleosomal DNA and across two neighboring regions. The frequency change rate for all dinucleotide types is illustrated in Figure 3. As shown in Figure 3A, significant flanking was observed surrounding the position of the central nucleosome on AA/TT, AT and TA, with a gentle bell shape around the central nucleosome for other dinucleotide types, such as C/G-based dinucleotides. Here, the frequency change rate can be treated as the preference of dinucleotides for nucleosomes, and according to the result shown in Figure 3A, AA/TT, AT and TA, which are also abundant in the nucleosome-free region before the TSS (Supplementary Fig. S7), can be regarded as nucleosome unfavorable dinucleotides. The flanking pattern appears to be reasonable because it shows that the position for a stable nucleosome has less disfavored dinucleotides in the central nucleosomal DNA region and more disfavored dinucleotides in the two outer sides of the nucleosomal region, just like a trap to capture or drop a nucleosome in the central region. This flanking pattern agrees with the definition of NCS because the nucleosome stability quantified from NCS is achieved by measuring the shift of the nucleosome center, and the disfavored dinucleotide types on both outer regions could theoretically limit the nucleosome from shifting. Moreover, we also applied NCS on one previous study of nucleosome positions (Jiang and Pugh, 2009), and NCS shows the similar flanking pattern and has a better result than the original study as illustrated in Supplementary Figure S8. To our knowledge, no existing study addresses the clear flanking pattern on dinucleotide preference for nucleosome stability until using NCS.

Based on nucleosome properties and the frequency pattern of different dinucleotide types described in the Supplementary Information, the nucleosomal region was partitioned into six regions, and the average frequency change rate for each dinucleotide and each region is shown in Figure 3B. We found that different contributions of each dinucleotide type to nucleosome stability and, for each dinucleotide, the extent of influence on different nucleosome regions were not the same. Except for AA/TT, AT and TA, which were dramatically reversed in the two neighboring regions, the patterns of frequency change rate mainly focused on the central region, and had only small values close to 0 in the regions of two neighboring nucleosomes (91–236 and 237–400 bp). This fact also implies the efficiency of NCS because the score scheme of NCS is for the central nucleosome, not for neighboring nucleosomes.

3.3 The negative relationship between NCS and poly(dA:dT)

By NCS, we observed significant patterns from A/T-based dinucleotides, and showed that AA/TT, AT and TA are disfavored dinucleotide types for nucleosomes, contrary to the result by NCP/noise. In fact, several previous studies have indicated that poly-A/T tracts, referred to as poly(dA:dT), tend to disfavor nucleosome formation (Field *et al.*, 2008; Segal and Widom, 2009a), associate with the nucleosome-free region (Wu and Li,

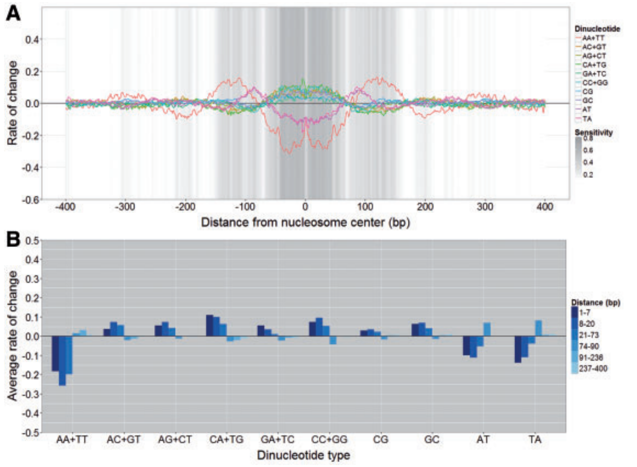

Fig. 3. (**A**) The frequency change rate (smoothing window size = 5 bp) surrounding the central nucleosome, illustrating the preference of each dinucleotide for a stable nucleosome. The sensitivity for each base pair quantifies the sensitive degree of dinucleotide to nucleosome score (NCS). (**B**) Average frequency change rate of different nucleosomal regions for each dinucleotide type. (A bigger figure is given in the Supplementary Data)

2010) and influence nucleosome organization and gene expression (Raveh-Sadka *et al.*, 2012). Therefore, we further showed the relationship between NCS and poly(dA:dT) by classifying the nucleosome data into 20 classes, based on ascending order of NCS, and observing the amount of poly(dA:dT) for each class. In Figure 4A, poly-A/T tracts with length ≥3 bp were detected to calculate the ratio of poly(dA:dT) on nucleosomal DNA regions. The poly(dA:dT) ratio decreased as the nucleosome stability increased, indicating poly(dA:dT) as disfavored sequences for nucleosomes, supporting previous results from the literature; the differences between the ratio of the lowest and highest classes are 7.24, 10.55 and 11.96% when the A/T switch becomes more tolerant, showing that the effect of poly-A/T tracts with one or two A/T switches is greater than that for pure poly(A) or poly(T) on nucleosome stability. Furthermore, considering the different sequence preferences of $(H3–H4)_2$ tetramer and H2A–H2B dimer, we divided the nucleosomal DNA region into inner and outer parts, and the extent of poly(dA:dT) influence on nucleosome stability was different. As illustrated in Figure 4B, the poly(dA:dT) ratio is more important for nucleosome stability in the central nucleosomal DNA region than in the outer region, stressing the effect of poly(dA:dT) on the central region of nucleosomal DNA. This phenomenon implies that the central nucleosome region mainly dominates its stability, consistent with the result in Section 3.1. Moreover, this phenomenon, under different A/T switches, can be observed in Figure 7, which shows that the difference between two patterns of poly(dA:dT) ratio separately from the central and outer regions becomes more clear when including an A/T switch. This observation suggests that the role of the AT or TA dinucleotide in the central region is more important than in the outer region, consistent with Supplementary Figure S9, which shows that the frequency of the AT/TA dinucleotide is more considerable in the central region than in the outer region.

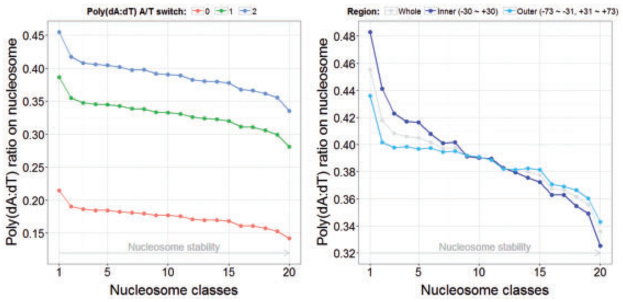

Fig. 4. Brogaard's data were classified into 20 classes based on the ascending order of NCS. (**A**) The ratio of poly(dA:dT) on nucleosomal DNA (the length of poly(dA:dT) ≥ 3 bp) and A/T switch means the time of poly(A) changing to poly(T) or poly(T) changing to poly(A) that poly(dA:dT) sequences can tolerate. (**B**) The poly(dA:dT) ratio with A/T switch = 2 on the central and outer nucleosomal regions

3.4 The opposite tendency between NCS and NCP/noise

Because we observed the opposite tendency on sequence preference between NCS and NCP/noise score from Brogaard *et al.*'s study, as stated in Section 3.1, we finally made a comparison between them and discussed this issue. In Brogaard *et al.*'s study, there are two kinds of scores: NCP measures the degree of cleavages observed in positions that conform to two chemical sites near the nucleosome center, and NCP/noise is a signal-to-noise ratio filtering out background noise and is the main nucleosome score used in their study (Brogaard *et al.*, 2012). Therefore, the frequency change rates for nucleosomal DNA under NCP/noise, NCP and noise are depicted in Figure 5. We will discuss the NCP/noise score first and leave discussion on the other two for the next section. According to the pattern of NCP/noise, the A/T-based dinucleotide preference in Figure 5A and poly(dA:dT) preference in Supplementary Figure S10 both show that strong nucleosomes favor A/T-based polynucleotides. Even though this observation is contrary to the result of NCS and previous studies, the expression of significant periodic preference for AA/TT in Figure 5A is not totally unreasonable because in addition to support disfavoring poly(dA:dT) in the literature, previous research found a preference for periodic dinucleotides (Wang *et al.*, 2008), and an AA/TT dinucleotide for every 10 bp makes DNA bend more favorably for nucleosomes (Prytkova *et al.*, 2011). These two opposite phenomena have also been indicated (Wal and Pugh, 2012) and may need more biological experiments to unveil the connection between them.

Although the results of NCS and NCP/noise are opposite on AA/TT/AT/TA dinucleotide preference, the relationship between these two score schemes is not simply negative because the Spearman correlation of NCS and NCP/noise is 0.339, and the distributions of nucleosome occupancy around TSS for NCS and NCP/noise are similar. As illustrated in Figure 6, the results of both score schemes support the barrier model, which means that the +1 nucleosome forms a barrier, causing a uniform positioning that gradually decays in nucleosome positioning afterward (Mavrich *et al.*, 2008). However, the pattern based on NCS more strongly supports the barrier model because the peaks of +1 to +6 nucleosome occupancy decrease step by step.

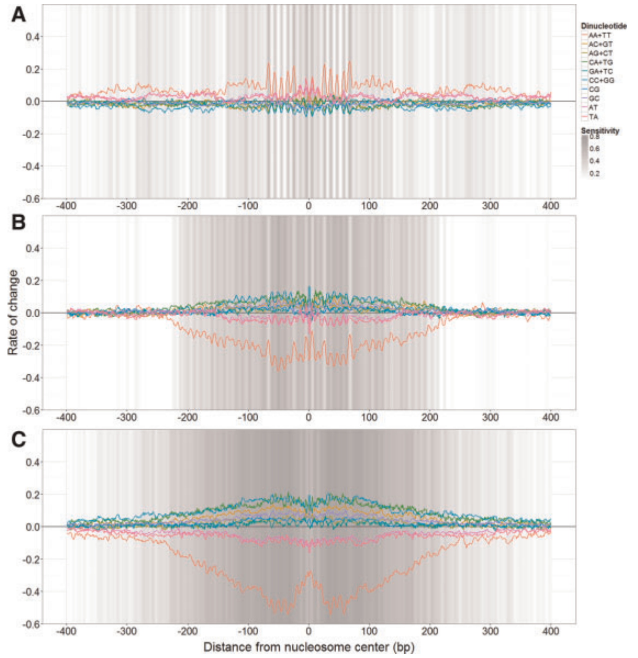

Fig. 5. The frequency change rate (smoothing window size = 5 bp) surrounding the central nucleosome, illustrating the preference of each dinucleotide for nucleosomes with a higher score in (**A**) NCP/noise (**B**), NCP and (**C**) noise

3.5 Further analysis on Brogaard *et al.*'s score

Because the score of NCP/noise is composed of NCP and noise, we further discussed these two score schemes and the interesting results that we found, represented in Figure 5B and C. First, the tendencies of AA/TT and AT/TA from the two schemes are similar, which is against the general expectation about the signal of noise, and means there could be sequence preference or bias in the noise score. According to the authors of Brogaard *et al.*'s study, low levels of non-specific chemical cleavages can be observed due to unknown reasons. Consequently, the noise was defined to account for the background cleavages and was estimated by the amount of cleavages in the neighboring region. This implies that the score of noise is not only a pure measure of background noise but could also include the nucleosome occupancy or strength to some extent. We think that is the reason why these two scores show similar preference for specific dinucleotides. Second, the widespread disfavoring of AA/TT and AT/TA is basically consistent with the result of NCS, and is similar to the results from our source positioning data for NCS in Supplementary Figure S11. This situation indicates that the tendency based on NCS is not completely opposite of what was observed in Brogaard *et al.*'s study, as the result of NCP/noise suggests.

We further checked the poly(dA:dT) ratio for all score schemes in Figure 7, and found that the value of noise is very sensitive to the poly(dA:dT) ratio (the turning point in Class 1 is possibly because the values of noise were all set to 0.5 when the original scores <0.5). Because the values of noise actually mix concepts of unknown background noise and nucleosome occupancy, we presume the high sensitivity to poly(dA:dT) is a result of these two factors, and that the unknown background noise may have bias on A/T-based tracts. Moreover, the four-template Poisson cluster model, which was used in Brogaard *et al.*'s study to calculate nucleosome scores, shows there could be some sequence preferences that affect the relative amount of primary and secondary chemical cleavage sites near the nucleosome center (Brogaard *et al.*, 2012).

According to the pattern of NCP and noise, and the slight turning point for Class 1 in Figure 7, we think that the score scheme of NCP was dominated to some extent by the score of noise. In Supplementary Figure S12, the average nucleosome occupancies around the TSS for NCP and noise show that the peak of the original +1 barrier nucleosome is dramatically pulled down compared with the +2 nucleosome, meaning that the pattern of NCP was influenced by the pattern of noise.

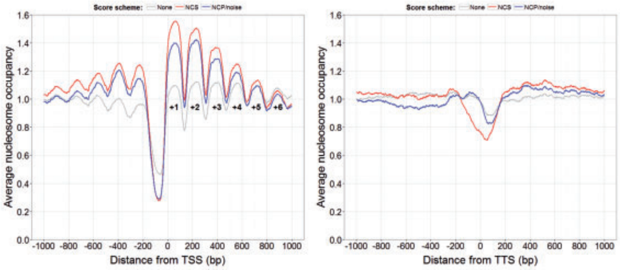

Fig. 6. The distribution of average nucleosome occupancies around the TSS and TTS for NCS and NCP/noise shows that the patterns under the two different score schemes are similar, but the NCS supports the model of the +1 barrier nucleosome after the TSS to a greater degree

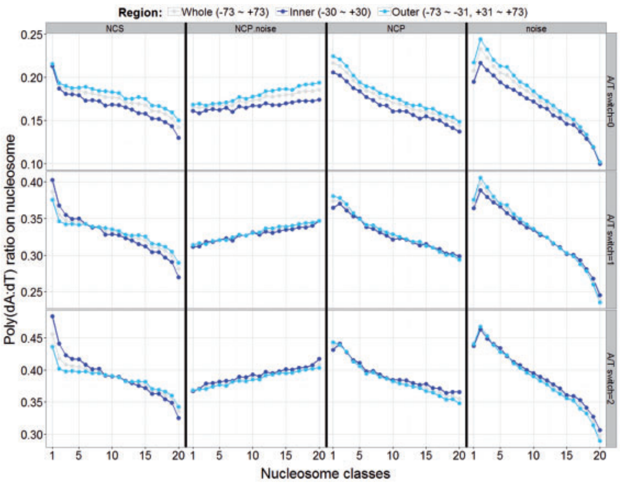

Fig. 7. The poly(dA:dT) ratio of nucleosomal DNA for different levels of nucleosome score, under four score schemes and different tolerating A/T switches on inner and outer nucleosomal DNA regions

4 CONCLUSION

The calculation of NCS is based on the incorporation of mass raw sequencing data, and it provides reliability in measuring nucleosome stability. This is because using a unifying processing method for all raw sequencing data can diminish the variance of each positioning data, and combining different PNs can supplement the insufficiency of single data. In this study, we applied NCS on the precise nucleosome positioning data from Brogaard *et al.*'s work. We then represented the efficiency of NCS based on known sequence characteristics of nucleosomal DNA, and showed the differences and possible limits of original Brogaard *et al.*'s scores at the same time.

Based on NCS, a clear dinucleotide preference for stable nucleosomes was illustrated, and significant flanking tendencies around nucleosomal DNA were also revealed for the dinucleotides AA/TT, AT and TA. The results of A/T-based dinucleotide and poly(dA:dT) show that the degrees of sequence preference for the central and outer regions of nucleosomal DNA are different, which is basically due to the fact that core histone $(H3\text{-}H4)_2$ tetramers dominate the free energy of histone–DNA interactions. In the comparison of NCS and Brogaard *et al.*'s score, the pattern of NCS showed more reasonable behaviors on the influence of central nucleosomal DNA regions, and disfavored the A/T tract across nucleosomal DNA, but we cannot judge or find definite reasons to explain the behaviors of the score from Brogaard *et al.*'s study. Nevertheless, some possible assumptions based on observations of sequence characteristics were provided in our study.

Because Brogaard *et al.*'s data support precise nucleosome positions and NCS reliably reflects the nucleosome stability, combing two data sources can enhance the future studies on epigenetic regulation or nucleosome stability, which is affected by several factors besides sequence preference. Moreover, we also provide the data of original NCS for each base pair in the Supplementary Data, so it will be easy to apply NCS to genome-wide-related data, such as data of histone variant, site of histone modification or even transcriptional factor binding site, to investigate the relationship with nucleosome stability.

ACKNOWLEDGEMENTS

The authors thank Darby Tien-Hao Chang, Chao-Hsuan Ke and Jiun-Huang Ju for their helpful comments.

Funding: This work was supported in part by National Science Council, Taiwan [NSC100-2627-B-006-011].

Conflict of Interest: None declared.

REFERENCES

Albert,I. *et al.* (2007) Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature*, **446**, 572–576.

Bohm,V. *et al.* (2011) Nucleosome accessibility governed by the dimer/tetramer interface. *Nucleic Acids Res.*, **39**, 3093–3102.

Brogaard,K. *et al.* (2012) A map of nucleosome positions in yeast at base-pair resolution. *Nature*, **486**, 496–501.

Chung,H.R. *et al.* (2010) The effect of micrococcal nuclease digestion on nucleosome positioning data. *PloS One*, **5**, e15754.

Cole,H.A. *et al.* (2012) Perfect and imperfect nucleosome positioning in yeast. *Biochimica et Biophysica Acta*, **1819**, 639–643.

Dreszer,T.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.

Field,Y. *et al.* (2008) Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.*, **4**, e1000216.

Jiang,C. and Pugh,B.F. (2009) A compiled and systematic reference map of nucleosome positions across the *Saccharomyces cerevisiae* genome. *Genome Biol.*, **10**, R109.

Kaplan,N. *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.

Kaplan,N. *et al.* (2010) Nucleosome sequence preferences influence in vivo nucleosome organization. *Nat. Struct. Mol. Biol.*, **17**, 918–920; author reply 920–922.

Kodama,Y. *et al.* (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.

Lee,W. *et al.* (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.

Li,B. *et al.* (2007) The role of chromatin during transcription. *Cell*, **128**, 707–719.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu,X. *et al.* (2006) Whole-genome comparison of Leu3 binding *in vitro* and *in vivo* reveals the importance of nucleosome occupancy in target site selection. *Genome Res.*, **16**, 1517–1528.

Mavrich,T.N. *et al.* (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, **18**, 1073–1083.

Prytkova,T.R. *et al.* (2011) Modeling DNA-bending in the nucleosome: role of AA periodicity. *J. Phys. Chem. B*, **115**, 8638–8644.

Raveh-Sadka,T. *et al.* (2012) Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.*, **44**, 743–750.

Rizzo,J.M. *et al.* (2011) Tup1 stabilizes promoter nucleosome positioning and occupancy at transcriptionally plastic genes. *Nucleic Acids Res.*, **39**, 8803–8819.

Segal,E. *et al.* (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.

Segal,E. and Widom,J. (2009a) Poly(dA:dT) tracts: major determinants of nucleosome organization, *Curr. Opin. Struct. Biol.*, **19**, 65–71.

Segal,E. and Widom,J. (2009b) What controls nucleosome positions? *Trends Genet.*, **25**, 335–343.

Thastrom,A. *et al.* (2004) Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. *J. Mol. Biol.*, **338**, 695–709.

Tsankov,A.M. *et al.* (2010) The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol.*, **8**, e1000414.

Tsui,K. *et al.* (2011) Evolution of nucleosome occupancy: conservation of global properties and divergence of gene-specific patterns. *Mol. Cell. Biol.*, **31**, 4348–4355.

van der Heijden,T. *et al.* (2012) Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. *Proc. Natl Acad. Sci. USA*, **109**, E2514–E2522.

Wal,M. and Pugh,B.F. (2012) Genome-wide mapping of nucleosome positions in yeast using high-resolution MNase ChIP-Seq. *Methods Enzymol.*, **513**, 233–250.

Wang,J.P. *et al.* (2008) Preferentially quantized linker DNA lengths in *Saccharomyces cerevisiae*. *PLoS Comput. Biol.*, **4**, e1000175.

Weiner,A. *et al.* (2010) High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res.*, **20**, 90–100.

Whitehouse,I. *et al.* (2007) Chromatin remodelling at promoters suppresses antisense transcription. *Nature*, **450**, 1031–1035.

Wu,R. and Li,H. (2010) Positioned and G/C-capped poly(dA:dT) tracts associate with the centers of nucleosome-free regions in yeast promoters. *Genome Res.*, **20**, 473–484.

Xi,Y. *et al.* (2011) Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation. *Genome Res.*, **21**, 718–724.

Xu,Z. *et al.* (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature*, **457**, 1033–1037.

Yuan,G.C. *et al.* (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*, **309**, 626–630.

Zhang,Y. *et al.* (2008) Identifying Positioned Nucleosomes with Epigenetic Marks in Human from ChIP-Seq. *BMC Genomics*, **9**, 537.

Zhang,Y. *et al.* (2009) Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat Struct Mol Biol*, **16**, 847–852.