EVALUATION OF DNA INTRAMOLECULAR INTERACTIONS FOR NUCLEOSOME POSITIONING IN YEAST

MICHAEL FERNANDEZ[1]
re0701m@bio.kyutech.ac.jp

SATOSHI FUJII[1]
s.fujii@bio.kyutech.ac.jp

HIDETOSHI KONO[2]
kono.hidetoshi@jaea.go.jp

AKINORI SARAI[1]
sarai@bio.kyutech.ac.jp

[1] *Department of Bioscience and Bioinformatics, Kyushu Institute of Technology (KIT), Iizuka, 820-8502, Japan.*
[2] *Computational Biology Group, Japan Atomic Energy Agency, Kyoto, 619-0215, Japan*

We calculated intramolecular interaction energies of DNA by threading DNA sequences around crystal structures of nucleosomes. The strength of the intramolecular energy oscillations at frequency ~10 bps for dinucleotides was in agreement with previous nucleosome models. The intramolecular energy calculated along yeast genome positively correlated with nucleosome positioning experimentally measured.

Keywords: chromatin, histone, indirect readout energy

1. Introduction

Nucleosomes are the primary organizational units of chromatin in *Eukaryotic* chromosomes. They are composed of about 147 DNA base pairs wrapped around a histone octamer [1] joined by linker DNA ranging from 20 to 80 bps [1]. Genome activities are largely regulated by nucleosomes together with the enzymes, which remodel and modify them. In this sense, nucleosome positioning can control the accessibility of underlying DNA to the nuclear environment. Translational and rotational settings define a nucleosomal midpoint relative to a given DNA locus and the orientation of DNA helix on the histone surface, respectively. Interaction of functional DNA sites with non-histone proteins influences sequence-directed nucleosome positioning playing an important functional role in determining the regularity of nucleosome location [2]. DNA regulatory elements may reside in linker regions between nucleosomes or along the nucleosome surface, where they may face inward (potentially inaccessible) or outward (potentially accessible). Nucleosome locations are partly defined by the underlying DNA sequence according to recent discoveries of nucleosome positioning sequences throughout the *S. cerevisiae* (yeast) genome [2]. AA/TT dinucleotides recurring in 10-bp intervals and in counterphase with GC dinucleotides had been hypothesized to generate a curved DNA structure that favours nucleosome formation [3].

Several methods have been proposed to computationally analyze nucleosome positioning [4]. Particular interest gained the contextual specificity of nucleosome after the first findings of DNA bending patterns related to periodic occurrences of certain

dinucleotides [4] and differences in Fourier transform analysis between nucleosomal and random DNA sequence [4]. A later work detected weak consensuses in nucleosome sites suggesting that no stringent conditions are imposed on a nucleotide sequence. However, periodicity of roughly 10 nucleotides was found in human exons and introns, and it was suggested that nucleosome position could be accessible by measuring 10-bp periodicities [4].

The imperfection and degeneracy of nucleosomal organization in the chromatin limits the application of classical computer analysis methods such as alignment and search for consensus [4]. The computational recognition of nucleosome positing sites is more difficult than transcription factor binding sites, which are characterized by consensuses and weighting matrices [5]. The typical complex coding responsible for a specific DNA conformation in histone complexes is rather generalizable and acceptable to many DNA sequences.

Structural analysis of a large number of protein-DNA complexes revealed a direct readout mechanism via contacts between amino acid residues and base-pairs, which are both redundant and flexible, suggesting that there is no simple code for the specificity of DNA-protein interactions [6,7]. In addition, protein-DNA binding specificity has been often modified by mutations of bases not in direct contact with amino acid residues, pointing out the importance of another indirect readout mechanism accounting for conformational changes (e.g. bending) [8] and/or flexibility [9] of DNA. These two mechanisms of recognition via direct protein-DNA contact and via DNA deformations, have been referred as intermolecular (direct) readout and intramolecular (indirect) readout, respectively. The specificity of intermolecular readout has been quantified based on the statistical analysis of the structures of protein-DNA complexes [9]. The energy of specific interactions between bases and amino acids for protein-DNA complexes have been calculated by empirical potential functions. Threading different DNA sequences on the protein-DNA framework and calculating the total energy have quantified differences in the fitness of various DNA sequences against the protein-DNA complex structure. This threading has enabled us to calculate Z-score against random sequences, as a measure of the specificity of the protein-DNA recognition, and to predict DNA target sites for regulatory proteins [9]. This framework was applied to the evaluation of DNA intramolecular interactions for nucleosome positioning in yeast.

2. Method and Results

2.1. Intramolecular Interaction Energy Calculation

Nucleosome positioning for a given DNA sequence is estimated by threading the constituent base pairs on the three-dimensional template constituted by the nucleosome core-particle structure, and calculating an intramolecular interaction energy in terms of the deviations of the base-pair step parameters that make up the structure from their

preferred equilibrium values [9]. A self-consistent component was added to the approach described by Olson et al. [10]. The conformation energies were approximated using a harmonic function:

$$E_{DNA}=1/2\Sigma\Sigma f_{ij}\Delta\Theta_i\Delta\Theta_j,$$

in which Θ_i represents the base-step parameters, and f_{ij} are the elastic force constants impeding deformation of the given base step $\Delta\Theta_i=\Theta_i - <\Theta_i>$, in which $<\Theta_i>$ is the average base-step parameter. The base-step parameters used were shift, slide, rise, tilt, roll, and twist. The definitions of these parameters are given as in the literature [11]. Note that we only gave the parameters for the ten mutually distinct base steps, while the remaining parameters were derived from symmetry relations [11]. The unknown parameters f_{ij} and $<\Theta_i>$ were determined by statistical analysis of non-redundant protein-DNA complexes [9]. Setting up a covariance matrix from observed distributions of Θ_i thus refers to an effective inverse harmonic force-constant matrix. Inversion of this matrix transformed it to a force-constant matrix in the original coordinate basis. The total intramolecular energy of a given complex structure was calculated as the sum of all the base steps.

2.2. Oscillation Pattern of Dinucleotides Along the Nucleosome Structure

A total of 15 nucleosome-containing crystal structures (histone similarity indexes lower than 30%) were retrieved from the Protein Data Bank (PDB). 3D coordinates and sequences of the nucleosome templates were extracted from the PDB files. We calculated the indirect readout energies for ten mutually distinct DNA dinucleotides at all the positions within the nucleosome structure, by sequence-structure threading over the nucleosome templates. Then, FFT was applied to evaluate the oscillation pattern of the indirect readout energies for all the dinucleodites (Fig. 1). The peaks are evident at ~34 degrees/base for some dinucleotides, corresponding to the turn of double-helix DNA. Furthermore, the magnitude depends on the dinucleotide sequence (Fig. 2): CT, CC and AA exhibit the highest peak values in descending order, whereas dinucleotide GC exhibits the lowest peak values. The peak amplitudes were normalized dividing by the average amplitude.

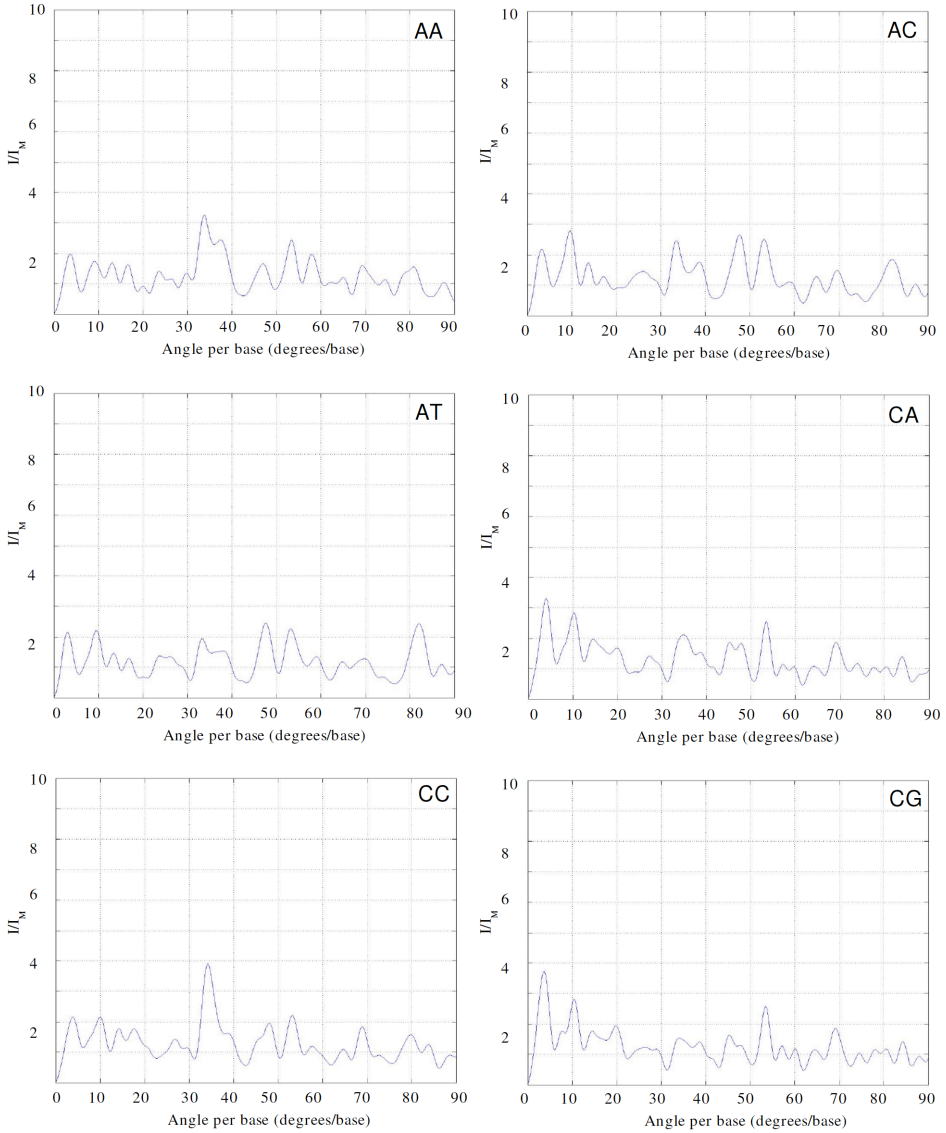

Figure 2: FFT average frequencies of indirect readout energies along nucleosome sequence for 10 dinucleotides. (Continue)

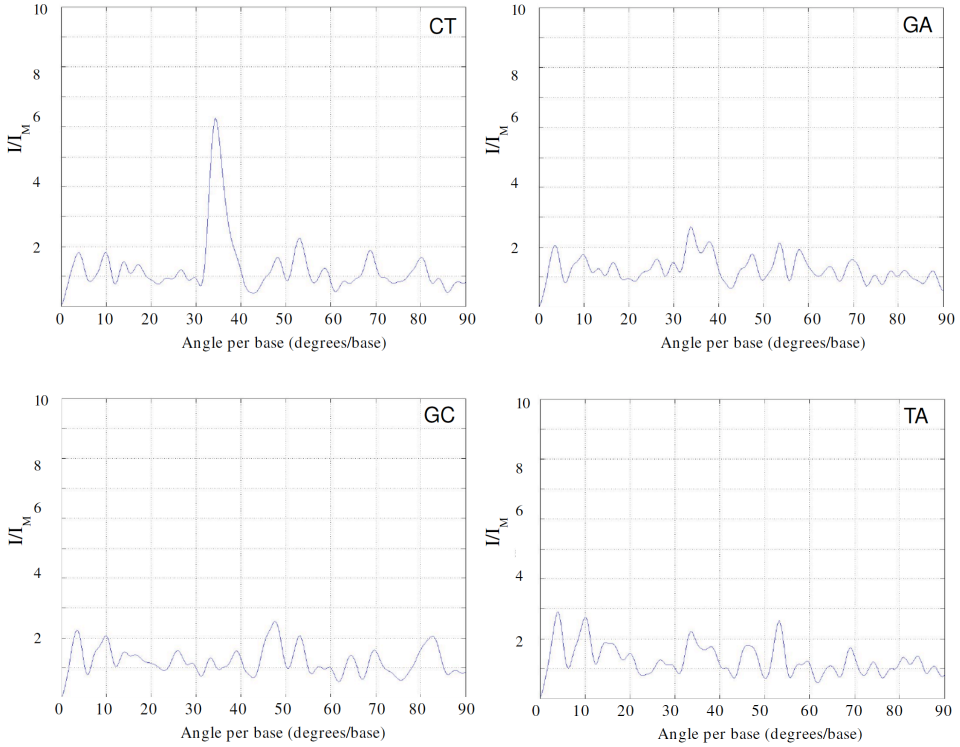

Figure 1: FFT average frequencies of indirect readout energies along nucleosome sequence for 10 dinucleotides.

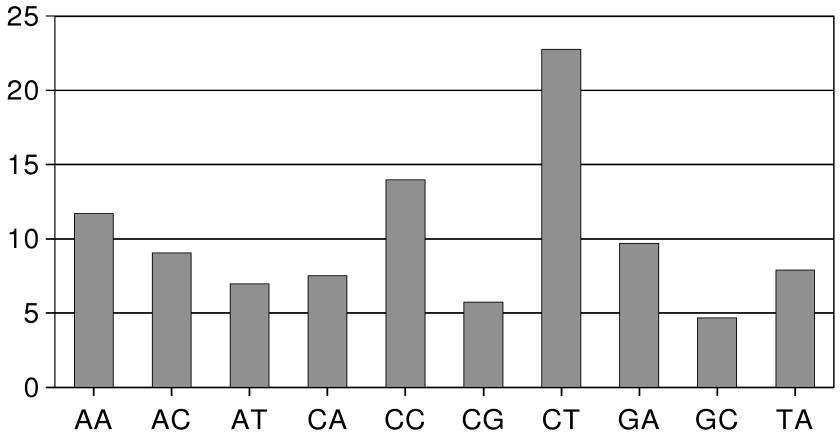

Figure 2: Relative intensity of FFT peaks at ~34 degrees/base of the indirect readout energies for the 10 dinucleotides.

2.3. Intramolecular Energy Profile of Yeast Genome

We used the DNA intramolecular energy to evaluate the nucleosome occupancy profile in yeast genome. In order to evaluate the probability that a nucleosome occupies a given position, we computed the energy by threading the corresponding DNA fragment against the nucleosome template. A six-parameter potential including twist, tilt, roll, shift, slide and rise dinucleotide step parameters of DNA double helix enabled us to compute the intramolecular interaction energy of DNA fragments of given sequence for the final superhelical structure around the histone core of crystal nucleosome structures used as templates.

The genome of *S. cerevisiae* was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/) and the experimental nucleosome occupancy data for *S. cerevisiae* (log2ratio) were retrieved from ref. 12). The intramolecular interaction energy landscape for nucleosome occupancy in yeast was obtained by computing Z-score values in windows centered at each nucleotide position along the yeast genome sequence using the best resolution template nucleosome (PDB code: 1KX5). The ability of the theoretical Z-scores to discriminate between nucleosome forming and inhibiting sequences was evaluated by receiver operating characteristic ROC analysis. Sub-sequences with extreme experimental nucleosome occupancies were labeled as nucleosome forming (high scores) and non-forming (low scores). For the whole dataset, the ROC scores were 0.7 and 0.8 for 10% and 1% of DNA residues having extreme score values, respectively.

On a large scale, when the nucleosome occupancy patterns of all promoters are averaged, the nucleosome-depleted region (NDR) is evident. Smoothing the energy landscape by a window of 4 bps, corresponding to the resolution of the experimental data, led to the energy profile, which was compared to *in vivo* nucleosome occupancy data as determined by measuring the accessibility to MNase of genomic DNA in chromatin. The average nucleosome intramolecular energy signature for yeast is clearly oriented at transcription start sites (TSSs). Precisely aligned nucleosome occupancy signal by TSSs and averaged all genes show higher values of intramolecular energy, in good agreement with the experimental scores (Figure 3).

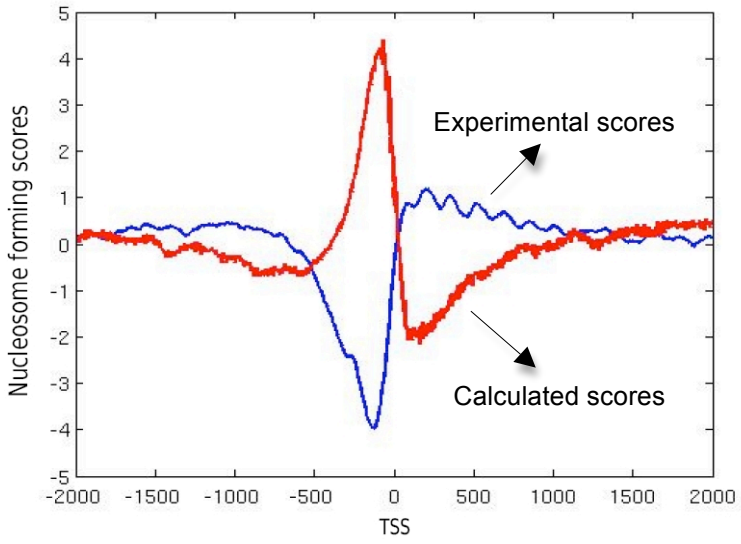

Figure 3. Nucleosome forming profile at yeast TSSs. Intramolecular interaction energy calculated in this study and experimental scores from ref. 12. Both scores are rescaled arbitrarily (intramolecular interaction energy on the right axis).

3. Discussions

Regularities of nucleosome positioning have been reported to depend on dinucleotide frequencies in the genome. The GC/AT-richness of a sequence has been considered as the strongest single factor among k-mer frequencies in determining its nucleosome formation potential in *S. cerevisiae* genome [13]. However, some author reported a species dependency in the nucleosome formation potential of genome. In this regards, regularity studies in human genome reflected that dominance of oscillating GG and CC dinucleotides in human nucleosomes and the contribution of AG(CT), GA(TC), and AA(TT) suggest a general nucleosome DNA sequence pattern-counterphase oscillation of RR and YY dinucleotides [14].

We have applied the intramolecular energy of DNA or indirect readout energy in protein-DNA recognition [9] to the problem of nucleosome positioning. The calculated intramolecular energy for the dinucleotides threaded along the nucleosome structure oscillates at frequency ~10 bps, in good agreement with previous nucleosome models which reported distinctive sequence motifs that recur periodically at the DNA helical repeat. However, the Fourier analysis showed that the relative strength of that peak in the power spectrum was highest for CT followed by CC and AA, and lowest for GC and CG, which is somewhat different from the result of sequence analysis. It remains to be seen how important is the phase information for the nucleosome positioning, and whether the

dinucleotides intermolecular energy is sufficient or not. We are currently investigating the relationship between the intramolecular energy profile and sequence features, and the role of periodicities or phasing in energy profile by including the effect of longer-range interactions.

We also compared the intramolecular interaction energy landscape to the nucleosome occupancy data in yeast [12]. We computed Z-score values based on the intramolecular energy, and the ROC analysis has shown that the ability of computed Z-scores to discriminate between nucleosome forming and inhibiting sequences is high. The energy landscape was also compared to the *in vivo* nucleosome occupancy data as determined by measuring the accessibility of MNase to genomic DNA in chromatin. The average nucleosome intramolecular energy signature for yeast is clearly oriented at TSSs, and the upstream NDR is evident. Precisely aligned nucleosome occupancy signal by TSSs and averaged all genes show higher values of intramolecular energy, in good agreement with the experimental scores, although the well-known ladder shape of nucleosomes positioned at downstream of TSSs is not well reproduced by the intramolecular interaction energy profile. However, the ladder might not be reflecting the intrinsic propensity of individual nucleosome formation. It should be noted that the nulceosome formation is a cooperative process regulated by chemical modifications, interactions with cofactors, solvent condition etc. We are currently working on a model that incorporates not only the intrinsic nucleosome positioning but also the effects of cooperativity.

References

[1] Luger,K., Mäder, A.W., Richmond, R.K., Sargent, D.F. and Richmond, T.J., Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389:251-260, 1997.

[2] Miele, V., Vaillant, C., d'Aubenton-Carafa, Y., Thermes, C., Grange, T., DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Research*, 36:3746–3756, 2008.

[3] Satchwell, S.C., Drew, H.R. and Travers, A.A., Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology*, 191:659–675, 1986.

[4] Levitsky, V.G., RECON: a program for prediction of nucleosome formation potential. *Nucleic Acids Research*, 32:W346–W349, 2004.

[5] Levitsky, V.G., Podkolodnaya, O.A., Kolchanov, N.A., Podkolodny, N.L., Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis. *Bioinformatics*, 17:998–1010, 2001.

[6] Matthews, B.W., Protein–DNA interaction. No code for recognition. Nature, 335:294–295, 1988.

[7] Pabo, C.O. and Nekludova, L., Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? *Journal of Molecular Biology* 301:597–624, 2000.

[8] Harrington, R.E. and Winicov, I., New concepts in protein–DNA recognition: sequence-directed DNA bending and flexibility. *Progress in Nucleic Acid Research and Molecular Biology*, 47:195–270, 1994.

[9] Gromiha, M.M., Siebers, J.G., Selvaraj, S., Kono, H., Sarai, A., Intermolecular and Intramolecular Readout Mechanisms in Protein–DNA Recognition. *Journal of Molecular Biology*, 337:285–294, 2004.

[10] Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M. and Zhurkin, V.B., DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proceedings of National Academy of Science USA*, 95:11163–11168, 1998.

[11] Olson, W.K., Bansal, M., Burley, S.K., Dickerson, R.E., Gerstein, M., Harvey, S.C., A standard reference frame for the description of nucleic acid base-pair geometry. *Journal of Molecular Biology* 313:229–237, 2001.

[12] Lee, W., Tillo, D., Bray, N., Morse, R.H., Davis, R.W., Hughes, T.R., Nislow, C., A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genetics*, 39:1235-1244, 2007.

[13] Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A.C., Field, Y., Moore, I.K., Wang, J.P.Z., Widom, V., A genomic code for nucleosome positioning. *Nature* 442:772–778, 2006.

[14] Kato, M., Onishi, Y., Wada-Kiyama, Y., Abe, T., Ikemura, T., Kogan, S., Bolshoy, A., Trifonov, E.N., Kiyama, R. Dinucleosome DNA of Human K562 Cells: Experimental and Computational Characterizations. *Journal of Molecular Biology*, 332:111-125, 2003.