In Silico Biology 6, 0008 (2006); ©2006, Bioinformation Systems e.V.
Workshop "GEOMETRY OF GENOME"
September 2005, Leicester, UK

DNA motifs and sequence periodicities

Lokesh Kumar, Matthias Futschik and Hanspeter Herzel*

Institute for Theoretical Biology, Humboldt University
Invalidenstr. 43, 10115 Berlin, Germany

* Corresponding author
Email: h.herzel@biologie.hu-berlin.de

Edited by E. Wingender; received October 30, 2005; revised January 26, 2006; accepted January 29, 2006; published February 11, 2006

Abstract

Genomic DNA sequences contain a wealth of information about the bendability and curvature of the DNA molecule. For example, the well-known 10-11 bp periodicities within genomes can be attributed to supercoiled structures or wrapping around nucleosomes. Such periodic signals have previously been examined mainly based on mono- or dinucleotide correlations. In this study, we generalize this approach and analyze correlation functions of longer motifs such as tetramers or poly(A) sequences. Periodically placed motifs may indicate regular protein binding or curvature signals. We detected various periodic signals e. g. strong 10-11 bp oscillations of periodically placed poly(A), poly(T) or poly(W) stretches. These observations lead to a new view on the intensively studied 10-11 bp periodicities.

Keywords: genomic DNA, periodicities, supercoiling, nucleosomes, motif, correlation, *C. elegans*, DNA bendability, poly(A) sequences

Introduction

Periodicities in DNA sequences have intensively been studied in the past decades. Already in 1980 Trifonov and Sussman found indications that 10-11 base-pair (bp) periodicities reflect DNA structure in chromatin. Zhurkin, 1981, pointed to the fact that -helices in proteins induce additionally DNA periodicities with a similar period (see Weiss and Herzel 1998 for a detailed discussion). These protein induced oscillations represent, however, only a minor fraction of the signal, since 10-11 bp periodicities have also been found in the third position of reading frames [Herzel *et al.*, 1998] and in non-coding DNA [Holste *et al.*, 2003; Dlakic *et al.*, 2004]. Using spectral analysis Widom, 1996, studied periodicities in eukaryotic genomes and found particularly strong signals in the DNA sequence of *C. elegans*. Interestingly, prokaryotic genomes also exhibit pronounced 10-11 bp periodicities associated with DNA supercoiling [Herzel *et al.*, 1998; Tomita *et al.*, 1999; Worning *et al.*, 2000]. With the aid of nonlinear curve-fitting [Herzel *et al.*, 1999] the specific periods of more than 100 genomes have been calculated

[Schieg and Herzel, 2004]. It turned out that genomes of archaea frequently display periods of around 10 bp associated with positive supercoiling [Herzel *et al*., 1998] whereas eubacteria exhibit periods between 10.7 and 11.5 bp reflecting negative supercoiling. Such observed 10-11 bp periodicities are commonly interpreted as bendability signals which support supercoiling or the wrapping of the DNA molecule around eukaryal or archaeal nucleosomes. This view is supported by multiple alignment of nucleosomal sequences [Satchwell *et al*., 1986; Ioshikkes *et al*., 1996], by artificial nucleosome positioning sequences [Shrader and Crothers, 1989] and by selecting sequences with high affinities for histone binding [Thastrom *et al*., 1999]. However, it should be noted that the observed periodicities are extremely weak. Typical amplitudes of correlation functions are in the order of 0.001, i. e. there is just a minor excess of appropriately spaced (di)nucleotides [Schieg and Herzel, 2004]. Only averaging over large genomic regions of about 100 kbp leads to detectable periodicities. Thus, the contribution of 10-11 bp periodicities to nucleosome positioning *in vivo* seems to be limited.

In this paper we explore the possibility that DNA periodicities reflect not only dinucleotide signals. As DNA curvature is governed by dinucleotides [Calladine and Drew 1987; Merino and Garciarrubio 2000], previous studies have been focused predominantly on the analysis of correlations or spectra of dinucleotides. In contrast, we analyse here longer oligonucleotides such as tetramers in order to detect periodically placed DNA motifs. DNA binding proteins might act as architectural elements to specify an optimal three-dimensional structure [Travers, 1990]. Even though most transcription factor binding sites are longer than 4 nucleotides, our search for tetramer signals might detect core motifs such as TATA or CAAT. We start with a systematic analysis of all 256 tetramers. It turns out that some motifs are indeed periodically placed along the DNA sequence with a variety of periods. The 10-11 bp periodicities are dominated by poly(A) and poly(T) stretches. This observation can give us a new perspective in the intensively studied field of DNA periodicities.

Methods

Correlation functions measure the excess of certain pairs of patterns at a distance of k base pairs. For the calculation of X-X autocorrelations, we count in the entire DNA sequence the number $N_{X\text{-}X}(k)$ of pairs of two identical patterns X and X separated by k base pairs. Here, X stands for a pattern such as X ='AATT'. Altogether there are $N - k - l + 1$ pairs in a sequence of length N, where l is the length of pattern. For example, the pattern AATT has the pattern length of 4 and thus the total number of pairs of this pattern (or any pattern of length 4) is $N - k - 3$ in a given sequence of length N. Consequently, the probability to find the pair X -X at the distance k can be estimated as:

$$P_{X\text{-}X} = N_{X\text{-}X}(k) / (N - k - l + 1)$$

The probability of a single pattern X is denoted by P_X. It is given by:

$$P_X(k) = N_X(k) / (N - k - l + 1)$$

In the above formula, N_X is the number of copies of the pattern in a sequence of length N. If the pairs at a distance k are statistically independent we find $P_{X\text{-}X}(k) = P_X(k) \cdot P_X(k)$. Thus the difference

$$C_{X\text{-}X}(k) = P_{X\text{-}X}(k) - P_X(k) \cdot P_X(k)$$

measures correlations at a distance of k base pairs. A positive peak of the covariance $C_{X\text{-}X}$ implies that there are more X-X pairs at a distance of k than expected by chance. Slow variations of the A + T content within genomes induce trends in most covariance functions. Consequently, most of the signals in our Figures 1-4 remain positive. In order to remove the trend we plot in Fig. 5 first order differences of the functions leading to oscillations around

zero.

All genomes were downloaded from NCBI (www.ncbi.nlm.nih.gov). Correlation functions were calculated with C programs available from the authors on request.

Results

Since periodicities of 10-11 bp are particularly strong in the genomic sequence of *C. elegans* [Widom, 1996; Schieg and Herzel, 2004] we start with a comprehensive analysis of tetramer periodicities in this organism. As described in the preceding section we calculate the covariance functions of all 256 tetramers. Due to strand symmetry [Lobry *et al.*, 1995] reverse complement tetramers such as AGAA and TTCT exhibit similar periodicities. We find a large variety of periodic signals including period 2, period 3 or period 8. Fig. 1 shows representative examples of such signals. Note that the amplitudes are fairly small, typically below 0.0001. The interpretation of period 3 is straight-forward: Due to a nonuniform codon usage [Sharp and Li, 1987; Holste *et al.*, 2000; Gorban *et al.*, 2003] the 3 positions in the reading frame have different compositions. This subsequently induces periodicities of longer oligonucleotides as well. For example, a relative high amount of nucleotide A in the second position of the reading frame implies an enhanced frequency of ATCA tetramers starting at the second position. Other periodicities are presumable due to repetitive sequences and will not be discussed in detail.

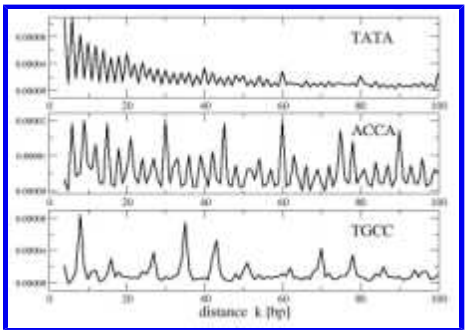


Figure 1: Autocorrelations of tetramer motifs in the genome of *C. elegans*. The upper graph shows period 2 of the TATA motif. In the middle graph a strong period 3 and a period 15 are visible whereas the lower graph exhibits 8 and 35 bp periodicities.

In the following, we focus on signals with periods around 10-11 bp since signals in phase with the helical period of DNA can induce curvature and affect bendability [Calladine and Drew, 1987]. Furthermore, periodically placed motifs might lead to a specific arrangement of DNA-binding proteins. In Figure 2 correlation functions with clear 10-11 bp periodicites are displayed. The amplitudes are somewhat higher than in Fig. 1. The tetramers with pronounced 10-11 bp periodicities have been compared with known transcription factor binding sites of *C. elegans* given in the Transfac database [Wingender *et al.*, 2000]. All consensus sequences of the *C. elegans* matrices were compared to the tetramers exhibiting clear 10-11 bp periodicities as in Fig. 2. However, we did not detected any clear similarities to core motifs of transcription factor binding sites.

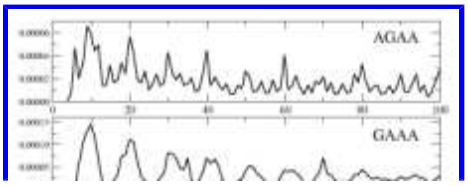
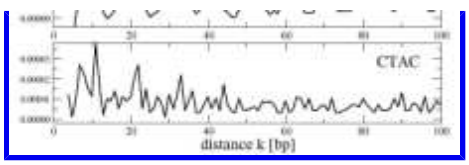


Figure 2: Autocorrelations of motifs AGAA, GAAA and CTAC in the genome of *C. elegans* displaying 10-11 bp periodicities.



Inspection of many correlation functions revealed that signals are particularly strong if the motifs are A+T rich. This is consistent with earlier studies in which correlations of the weakly binding nucleotides A and T have been analyzed as markers of 10-11 bp periodicities [Widom, 1996; Herzel *et al*., 1999]. Fig. 3 shows the oscillations of AAAA, TTTT, AAAT and WWWW autocorrelations. The signals are much stronger than all the other periodicities discussed above. An overrepresentation of poly (A), poly(T) and poly(W) stretches in genomic DNA is well known. Molecular mechanisms leading to such repetitions are the reverse transcription of poly(A)-tails in eukaryotic genomes and slippage of DNA-polymerases during replication. In the following we consider different lengths of poly(W) stretches. It turns out that the 10-11 bp periodicities look quite similar for stretches $(W)_n$ with n = 1, 2,..,8 (see Fig. 4). For even longer stretches the number of pairs in a certain distance becomes rather small and thus the signal-to-noise-ratio decreases. Similar periodicities as in Fig. 4 are visible in correlations of $(A)_n$ and $(T)_n$ motifs (n = 1, 2, ...8) albeit the signals are weaker. Figures 3 and 4 represent the main finding of this study. Periodically placed poly(A/T) stretches along the genomic sequence of *C. elegans* are a major source of the widely discussed 10-11 bp periodicities.

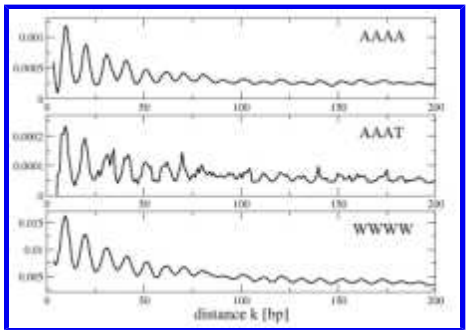


Figure 3: Poly(A/T) motifs are periodically placed along the genomic DNA of *C. elegans*. Autocorrelations of the motifs AAAA, AAAT and WWWW show pronounced 10-11 bp periodicities.

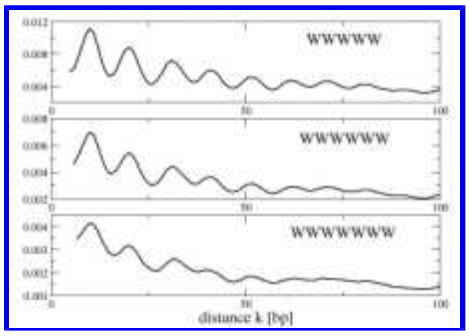


Figure 4: Autocorrelation functions reveal 10-11 bp periodicities of poly(W) motifs in the genome of *C. elegans*.

In order to test whether these results apply also to other genomes we compare in Fig. 5 WWWW tetramer oscillations in different genomes. In order to remove trends and to eliminate period 3 we show the first differences

of 3 bp running averages. As discussed above the signal is particularly strong for *C. elegans* (note the different scales). The oscillations in the other 3 genomes suggest that our findings seem to be relevant to other organisms as well.

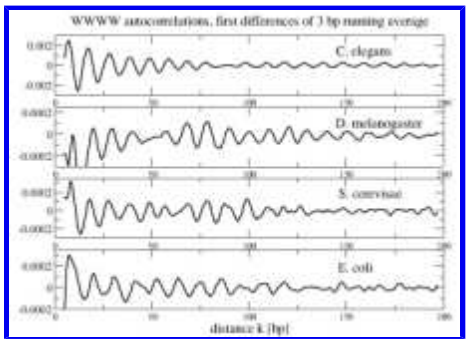


Figure 5: WWWW autocorrelations in the complete genomes of *C. elegans*, *D. melanogaster*, *S. cerevisiae* and *E. coli*.

Discussion

In earlier studies mono- and dinucleotide periodicities have been analysed in detail. In this paper we focused on periodically placed motifs. A comprehensive scan of all 256 tetramer-signals for the complete genome of *C. elegans* was carried out. Observed periodicities could not be traced back easily to core motifs of transcription factor binding sites. Very pronounced 10-11 bp periodicities have been observed for correlations of poly(A), poly(T) and poly(W) stretches, i. e. these motifs are found to prefer distances of multiples of the helical period. Some transcription factors such as TBP, SRF or the *C. elegans* factors Skn-1, DAF-16 or unc-86 bind to stretches of A/T nucleotides. Thus we cannot exclude an association of periodic protein binding sites and our observed oscillations.

In our view there is no straightforward interpretation of periodically placed poly(A/T) stretches. One might argue that dinucleotide correlations simply induce periodicities of longer oligonucleotides. However, in this case the amplitude of the oscillations should drastically shrink with the length of the motif as discussed in Schieg and Herzel, 2004. For instance, the amplitude of dinucleotide correlations induced by mononucleotide correlations should be 8 times smaller. The persisting relatively large amplitudes of the oscillations in Fig. 4 indicate that the correlations of poly(A/T) stretches constitute the primary signal that in turn lead to dinucleotide correlations described in earlier studies.

The strength of the poly(A/T)-signals in Figures 3 and 4 is related to the overrepresentation of poly(A/T) stretches in eukaryotic genomes. For example, the motif lexicon [Dyer *et al.*, 2004] indicates that AAAAAA hexamers in intergenic regions of *C. elegans* are 8 times more frequent than expected by chance (see http://genomics.wheatoncollege.edu/cgi-bin/lexicon.exe). The mechanisms leading to poly(A/T) sequences are well-known (reverse transcription of poly(A) tails, slippage of polymerases). It is, however, an open question how these processes relate to the observed 10-11 bp periodicities. The excess of poly(A/T) sequences is particularly strong in non-coding DNA. Interestingly, Dlakic *et al.*, 2004, found that 10-11 bp periodicities are much stronger in non-coding DNA as well.

Even though DNA curvature and bendability models are typically based on dinucleotides [Calladine and Drew 1987; Goodsell and Dickerson, 1994] it is known that poly(A) sequences induce curvature (see e. g. the review by Olson and Zhurkin, 1996). Electrophoretic mobility studies revealed that A tracts in phase with the helical period induce large curvature [Haran *et al.*, 1994]. Thus periodically placed poly(A/T) stretches found in this paper might support the optimal topology of genomic DNA.

Acknowledgements

We thank Szymon Kielbasa and Steven A. Brown for stimulating discussions regarding transcription factor binding sites and periodically arranged proteins as a putative source of 10-11 bp periodicities and two anonymous referees for helpful comments. Support was provided by Deutsche Forschungsgemeinschaft (SFB 618).

References

- Calladine, C. R. and Drew, H. R. (1997). Understanding DNA. Second Edition, Academic Press, London.
- Dlakic, M., Ussery, D. W. and Brunak, S. (2004). DNA Bendability and Nucleosome Positioning in Transcriptional Regulation. *In:* DNA Conformation and Transcription, Takashi Ohyaman (ed.), Eurekah Bioscience Database, chapter 14, pp.1-14.
- Dyer, B. D., LeBlanc, M. D., Benz, S., Cahalan, P., Donorfio, B., Sagui, P., Villa, A. and Williams, G. (2004). A DNA motif lexicon: cataloguing and annotating sequences. In Silico Biol. **4**, 0039.
- Goodsell, D. S. and Dickerson, R. E. (1994). Bending and curvature calculations in B-DNA. Nucleic Acids Res. **22**, 5497-5503.
- Gorban, A., Zinovyev, A. Y. and Popova, T. G. (2003). Seven clusters in genomic triplet distributions. In Silico Biol. **3**, 0039.
- Haran, T. E., Kahn, J. D. and Crothers, D. M. (1994). Sequence elements responsible for DNA curvature. Mol. Biol. **244**, 135-143.
- Herzel, H., Weiss, O. and Trifonov, E. N. (1998). Sequence periodicity in complete genomes of archaea suggests positive supercoiling. J. Biomol. Struct. Dyn. **16**, 341-345.
- Herzel, H., Weiss, O. and Trifonov, E. N. (1999). 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics **15**, 187-193.
- Holste, D., Grosse, I., Buldyrev, S. V., Stanley, H. E. and Herzel, H. (2000). Optimization of coding potentials using positional dependence of nucleotide frequencies. J. Theor. Biol. **206**, 525-537.
- Holste, D., Beirer, S., Schieg, P., Grosse, I. and Herzel, H. (2003). Repeats and correlations in human DNA sequences. Phys. Rev. E **67**, 061913.
- Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M. and Trifonov, E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. J. Mol. Biol. **262**, 129-139.
- Lobry, J. R. (1995). Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. **40**, 326-330.
- Merino, E. and Garciarrubio, A. (2000). The global intrinsic curvature of archael and eubacterial genomes is mostly contained in their dinucleotide composition and is probably not an adaptation. Nucleic Acids Res. **28**, 2431-2438.
- Olson, W. K. and Zhurkin, V. B. (1996). Twenty Years of DNA Bending. *In:* Biological Structure and Dynamics, R. H. Sarma and M. H. Sarma (eds.), Adenine Press, pp. 341-370.
- Satchwell, S. C., Drew, H. R. and Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. J. Mol. Biol. **191**, 659-675.
- Schieg, P. and Herzel, H. (2004). Periodicities of 10-11 bp as indicators of the supercoiled state of genomic DNA. J. Mol. Biol. **343**, 891-901.

- Sharp, P. M. and Li, W.-H. (1987). The codon Adaptation Index - a measure of directional synonymous codon usage bias, and its potential application. Nucleic Acids Res. **15**, 1281-1295.
- Shrader, T. E. and Crothers, D. M. (1989). Artificial nucleosome positioning sequences. Proc. Natl. Acad. Sci. USA **86**, 7418-7422.
- Thastrom, A., Lowary, P. T., Widlund, H. R., Cao, H., Kubista, M. and Widom, J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. J. Mol. Biol. **288**, 213-229.
- Tomita, M., Wada, M. and Kawashima, Y. (1999). ApA dinucleotide periodicity in prokaryote, eukaryote, and organelle genomes. J. Mol. Evol. **49**, 182-192.
- Travers, A. A. (1990). Why bend DNA? Cell **60**, 177-180.
- Trifonov, E. N. and Sussman, J. L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc. Natl. Acad. Sci. USA **77**, 3816-3820.
- Weiss, O. and Herzel, H. (1998). Correlations in protein sequences and property codes. J. theor. Biol. **190**, 341-353.
- Widom, J. (1996). Short-range order in two eukaryotic genomes: relation to chromosome structure. J. Mol. Biol. **259**, 579-588.
- Wingender, E., Chen, X., Hehl, R. Karas, H., Liebich, I., Matys, V., Meinhart, T., Prüß, M., Reuter, I. and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. **28**, 316-319.
- Worning, P., Jensen, L. J., Nelson, K. E., Brunak, S. and Ussery, D. W. (2000). Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*. Nucleic Acids Res. **28**, 706-709.
- Zhurkin, V. B. (1981). Periodicity in DNA primary structure is defined by secondary structure of the coded protein. Nucleic Acids Res. **9**, 1963-1971.