Genomic Data Analysis Using DNA Structure: An Analysis of Conserved Nongenic Sequences and Ultraconserved Elements

Eleanor J. Gardiner,*[,†,‡] Linda Hirons,[†,‡] Christopher A. Hunter,[†] and Peter Willett[‡]

Centre for Chemical Biology, Krebs Institute for Biomolecular Science, Department of Chemistry, University of Sheffield, Sheffield S3 7HF, United Kingdom, and Department of Information Studies, University of Sheffield, Sheffield S1 4DP, United Kingdom

Received September 9, 2005

Recent comparative studies of the human and mouse genomes have revealed sets of conserved nongenic sequences (CNGs) and sets of ultraconserved elements (UCEs). Both sets of sequences, which exhibit extremely high levels of conservation, extend over hundreds of bases and have no known function. Since there is no detectable sequence homology between paralogous CNGs or UCEs in either of the species, an alignment-free technique is needed for their analysis. We have previously compiled a database of the structural properties of all 32 896 unique DNA octamers, including information on stability, the minimum energy conformation, and flexibility. We have used Fourier techniques to analyze the UCEs and CNGs in terms of their octamer structural properties, to reveal structural correlations which may indicate possible functions for some of these sequences.

INTRODUCTION

Only 1−2% of the DNA of the human genome codes for proteins. Much of the remainder may be "junk", but comparative genomics suggests that a significant amount must serve some purpose. For example, recent comparisons between the mouse and human genomes found that 5% of the genome is conserved between the two species.[1] This means that, in addition to the protein-coding regions, about 3% of the genome is likely to be under evolutionary selection for some as yet unknown function. Human and mouse comparative studies have revealed sets of conserved nongenic sequences (CNGs)[2,3] and sets of ultraconserved elements (UCEs),[4] which exhibit extremely high levels of conservation, which extend over hundreds of bases and have no known function.

The CNGs were found in a comparative analysis of human chromosome 21 with the mouse genome. The CNGs are hundreds of bases long and have at least 70% ungapped identity between human and mouse in all cases and, thus, are much more highly conserved than protein-coding genes. More than a quarter of the CNGs on human chromosome 21 have been found in at least 10 other species. However, there is no significant sequence homology between pairs of CNGs within a species.[5] An extrapolation to the entire genome predicts that between 0.3% and 1% of the genome may be composed of such CNGs.[3] Triplication of a subset of the CNGs of chromosome 21 may contribute toward the Down's syndrome phenotype.[6]

UCEs are also hundreds of bases long and are absolutely conserved, between human and mouse, without gaps. Over half of the UCEs identified have no overlap with any exonic sequences.[4] Interestingly, UCEs were found on all chromosomes except for 21 and Y. The extraordinarily high degree of conservation of these elements strongly suggests their significance and underlines the need to develop new approaches to understanding the function of noncoding sequences. One hypothesis is that their function could be related to protein binding, as part of a system of DNA repair, gene expression, replication, packaging, or scaffold attachment.[7] A second hypothesis is that such highly conserved noncoding elements have only been found in vertebrates[8] and so may be important for vertebrate development. One factor that governs protein−DNA interactions and is expressed over length scales of hundreds of bases is the structural nature of the DNA. The fact that the conservation of both the CNGs and the UCEs between species is ungapped suggests a possible structural reason for the conservation, since the insertion of a single base can have a large effect on the DNA three-dimensional structure. A reasonable proposition is that the identification of common DNA structural motifs could allow annotation of the functional properties of noncoding parts of the genome and provide insights into likely protein partners or structural function.

In previous work, we constructed the potential energy surfaces for all unique tetramers, hexamers, and octamers in double helical DNA, as a function of the two principal degrees of freedom, slide and shift at the central step. The method is based on an ab initio treatment of the base stacking interactions and an empirical model for the backbone and environment.[9−14] From these potential energy maps, we compiled a database of the structural properties of all 32 896 unique DNA octamer sequences, including information on stability, the minimum energy conformation, and flexibility.[15,16] The most important properties are summarized in Table 1.

We demonstrated that DNA structural space is less diverse than DNA sequence space since many dissimilar octamers

* Corresponding author. Tel.: (+44) 0114 2222674; fax: (+44) 0114 278 0300; e-mail: e.gardiner@sheffield.ac.uk.
† Department of Chemistry, University of Sheffield.
‡ Department of Information Studies, University of Sheffield.

10.1021/ci050384i CCC: $33.50
Published on Web 01/06/2006

Table 1. Octamer Structural Properties

property	
three-step roll, three-step twist	roll (twist) of the central three steps of an octamer, considered as one giant step
flexibility	ability to increase or decrease roll or twist
three-step flexibility	ability to increase or decrease three-step roll or three-step twist

may possess similar structures.[16] Thus, although neither the CNGs or UCEs exhibit any sequence similarity within the human genome, we may be able to find common structural elements that might help elucidate possible commonality of function. These properties are described in detail by Gardiner et al.[15]

Fourier transforms may be used to isolate a signal from noisy data. Figure 1a shows the three-step roll of a nucleosome-wrapping DNA sequence measured from the crystal structure.[17] The power spectrum of the Fourier transform (Figure 1b) shows a strong peak at approximately 10.2 base pairs (bp) which corresponds to the helical repeat of DNA. Periodic roll values that are in phase with the helical repeat lead to a persistent bend in the DNA, and this is obviously what is required to spool it around the nucleosome. Fourier techniques can also be used to isolate patterns in DNA sequences, such as the presence of protein-coding regions.[18,19] Recently, for example, Sharma et al. used a discrete Fourier transform as the basis for their Spectral Repeat Finder, which isolates repetitive DNA sequence elements such as microsatellites.[20]

In this work we apply Fourier techniques to analyze the UCEs and CNGs in terms of their octamer structural properties, to reveal long-range structural correlations which may indicate possible functions for some of these sequences.

METHODS

For a given structural parameter (e.g., three-step roll from Table 1) and DNA sequence of length *N*, we proceed as follows: (1) Construct a vector composed of the value of the parameter for each of N-7 overlapping octamer. (2) Pad the vector with zeroes at the end if necessary, to give a vector whose length is a power of 2. (3) Take the Fourier transform of this vector. (4) Obtain its power spectrum.

To obtain an estimate of the noise, (5) shuffle the DNA sequence (i.e., the nucleotides). (6) Repeat steps 1−4 above to obtain the power spectrum of the shuffled DNA.

To compare the spectra with those found using a CNG or UCE sequence rather than structure, (7) construct a vector by replacing each A nucleotide with 1 and the remaining bases with 0. (8) Repeat steps 2−4 above to obtain the power spectrum of this vector. (9) Repeat steps 7−8 above for all C, G, T nucleotides in turn. (10) Sum the four spectra to give a total spectrum for the sequence. To avoid confusion, we refer to this as an *occurrence* spectrum, since it reflects the periodic occurrence of nucleotide types.

The Fourier transforms were performed using Matlab (www.mathworks.com).

To view the results for the CNGs or UCEs, we plot the mean of their spectra. A particular peak in a power spectrum indicates that a corresponding periodicity is present somewhere in the vector of structural property values under transformation. An advantage of our Fourier technique is, therefore, that the sequences do not need to be aligned in order for similarity in structural properties to be apparent.

We applied this method to the set of CNGs and the set of UCEs described in the Introduction. As we are interested in long-range interactions, we consider only those CNGs longer than 150 bp.

The CNGs[3] and UCEs[4] have been classified by the authors as belonging to subsets as detailed in Table 2. CNGs are described as being of *known* or *unknown* function. Known CNGs (kCNGs) overlap with regions such as pseudo genes, pseudo coding regions. Thus, although these kCNGs have no known function, they resemble known functional elements. Unknown CNGs (uCNGs) have not been associated with any functional elements. An UCE is classified as exonic (eUCEs) if any part of the UCE overlaps the mRNA of a known human protein-coding gene. Nonexonic UCEs (nUCEs), in contrast, show no evidence of transcription in any species. The remaining UCEs are classified as possibly exonic (pUCE). Table 2 also gives the base composition for each type of the CNGs and UCEs. The UCEs are very similar in base composition to the uCNG, having relatively more A's and T's and fewer C's and G's, whereas the kCNGs are different, with approximately equal amounts of all bases.

We also compared the mean values of the structural properties of the CNGs and UCEs with both the mean values from our octamer database and the mean values from human chromosome 21, which was considered to be representative of human genomic data (results not shown). We found no significant differences between the values for either the CNGs or the UCEs and those of the database or chromosome 21.

RESULTS AND DISCUSSION

CNGs. Figure 2 shows the mean occurrence power spectra of the CNG. Occurrence power spectra for the kCNGs (Figure 2a) show a pronounced peak at 3 bp. Such peaks are well-known to occur in the Fourier spectra of exonic sequences[18,21] and correspond to codon-containing DNA. As expected, the uCNGs do not have a 3 bp peak (Figure 2b), since they do not contain codons. Apart from this peak, the spectra in Figure 2 are generally featureless, merely decreasing in intensity somewhat from left to right. In both cases, the random spectra show no features at all but simply consist of a straight line with small fluctuations.

Figure 3 shows the mean spectra of the kCNGs for a selection of structural parameters. In each case, the mean randomized spectrum is also shown. In contrast to the occurrence spectra of Figure 2, each structural property has a distinctive shape to its power spectrum. For example, decreasing twist flexibility has a pronounced hump between 5 and 3 bp (Figure 3a), while the three-step roll is a sine wave (Figure 3h). This is true for the randomized kCNGs as well as genomic kCNGs, since the six-step overlap between neighboring octamers enforces a relationship between consecutive values of any structural parameter. The kCNG power spectra for most of the structural properties contain a pronounced peak at 3 bp (Figure 3a−e,g), demonstrating that the structural parameters can isolate known features. The only exceptions are three-step roll and three-step twist, which clearly have no such peak (Figure 3f,h).

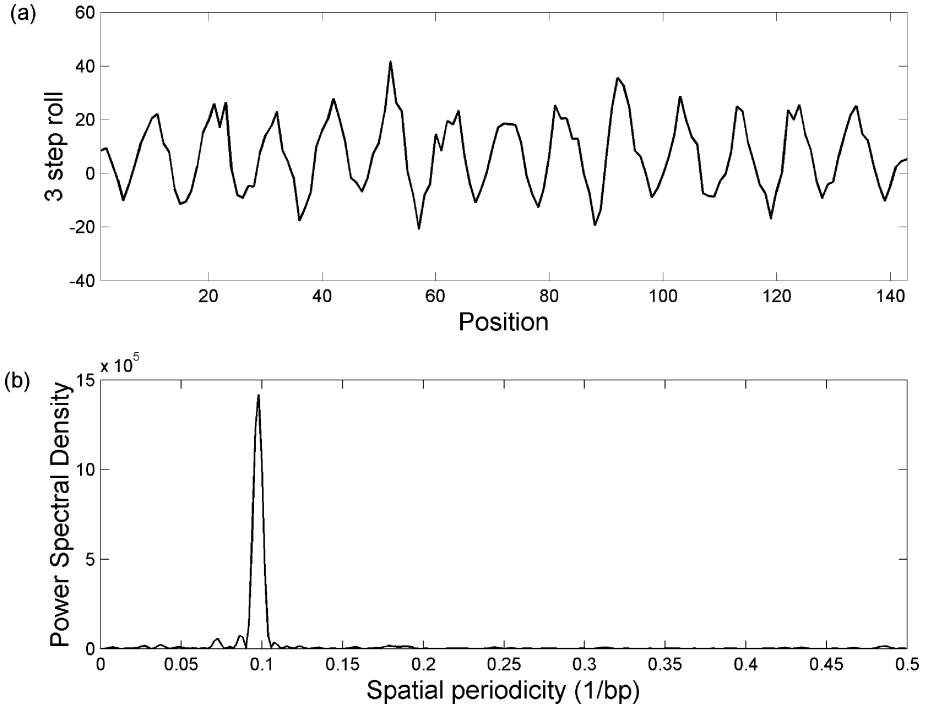

Figure 1. Nucleosome three-step roll. (a) Three-step roll, measured from the crystal structure of the nucleosome core particle, complexed with DNA.[17] (b) Power spectrum.

Table 2. Sequence Data

	function	number of sequences	mean sequence length	%A/T	%C/G
		CNGs			
kCNG	known	755	241	50.1	49.9
uCNG	unknown	869	205	62.1	37.9
		UCEs			
eUCE	exonic	111	254	58.0	42.0
pUCE	possibly exonic	114	268	60.9	39.1
nUCE	nonexonic	256	263	63.5	36.5

In Figure 4, we compare the power spectra of the kCNGs and uCNGs for a selection of structural properties, chosen because they contain differences between two sets which may be of importance. The uCNGs contain a pronounced peak in decreasing three-step-twist flexibility at ~22.2 bp which is not present in the kCNGs (Figure 4a). This peak is also present, although much less pronounced, in minor groove width and decreasing twist flexibility (Figure 4b,c). The value 22.2 bp is approximately twice the double-helical pitch of DNA and may indicate a structural feature of these CNGs.

For most properties (including the great majority of those not shown), the kCNGs show a slightly higher overall intensity, whereas for minor groove width and three-step roll (Figure 4b,d), the uCNGs have regions of higher-power spectral density.

UCEs. As for the kCNGs, the eUCEs show peaks at 3 bp in the flexibility power spectra (not shown), again indicating the presence of codons in the exonic UCEs.

Differences between the power spectrum and the randomized spectrum can be used to emphasize any signal present in the spectrum. Figure 5 shows the mean power spectra of the nUCE for three-step twist and three-step roll (Figure 5a,b). The difference spectra is also plotted (Figure 5c,d). These plots bring to the fore features, such as the peak at 12.4 bp in three-step twist, which are not obvious from the original plot. It is striking that this periodicity of 12.4 bp apparent in three-step twist, is twice that exhibited by three-step roll. These peaks are not present for the CNGs, nor for the eUCEs or pUCEs.

The presence of a peak in the mean power spectra does not necessarily mean that the peak is present in all of the individual UCEs. Figure 6a shows the mean power spectra for those 14 nUCEs whose peak in the three-step roll spectrum at 6.2 bp is greater than the mean plus two standard deviations of the peak value at this position over all nUCEs. Figure 6b shows the mean spectra for the remaining nUCEs. It is clear that this peak is a large feature for a subset of the nUCEs while being absent from many of the remainder. Figure 7a shows individual power spectra for 3 of the 14 nUCEs which possess the three-step roll periodicity, while Figure 7b shows individual power spectra for three of the nUCEs which do not exhibit this property.

Further Analysis of UCEs. Our analysis of the nUCEs has shown that at least 14 elements have a periodicity of 6.2 bp. These nUCEs are listed in Table 3. The sequence identifiers are taken from Bejerano et al.[4]

We next attempted to isolate the parts of the nUCE which contain the periodicity. We are currently developing methods which will automatically align sets of sequences using any common structural properties. In the current situation, we know the property of interest is a periodicity of 6.2 bp in three-step roll. We, therefore, generated an ideal three-step roll of period 6.2 bp, mean 7.4°, and amplitude 4.0° using the equation

$$\text{roll3} = 4.0 \sin[(i-1)2.0\pi/6.2] + 7.4 \qquad i = 1{:}60$$

The value 7.4° is the mean of the minimum energy three-step roll values of the nUCE, and 4.0° is the standard deviation. The method here is merely illustrative of the type

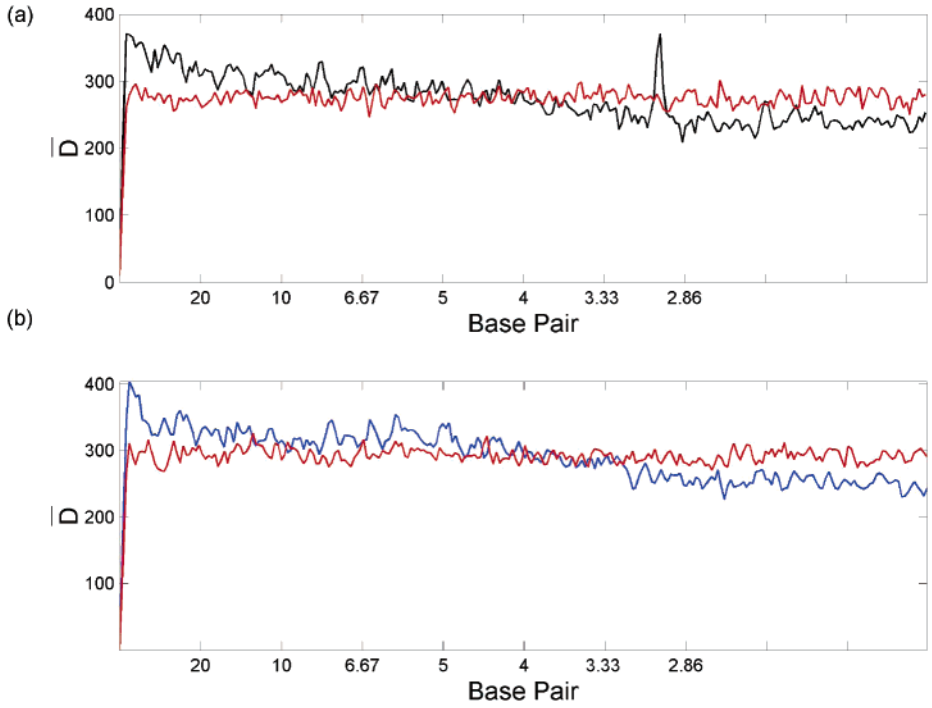

Figure 2. Occurrence power spectra of kCNG. $\bar{D}$ is the mean occurrence power spectra; red line, spectra of randomized sequences. (a) Black line, kCNG; (b) blue line, uCNG.

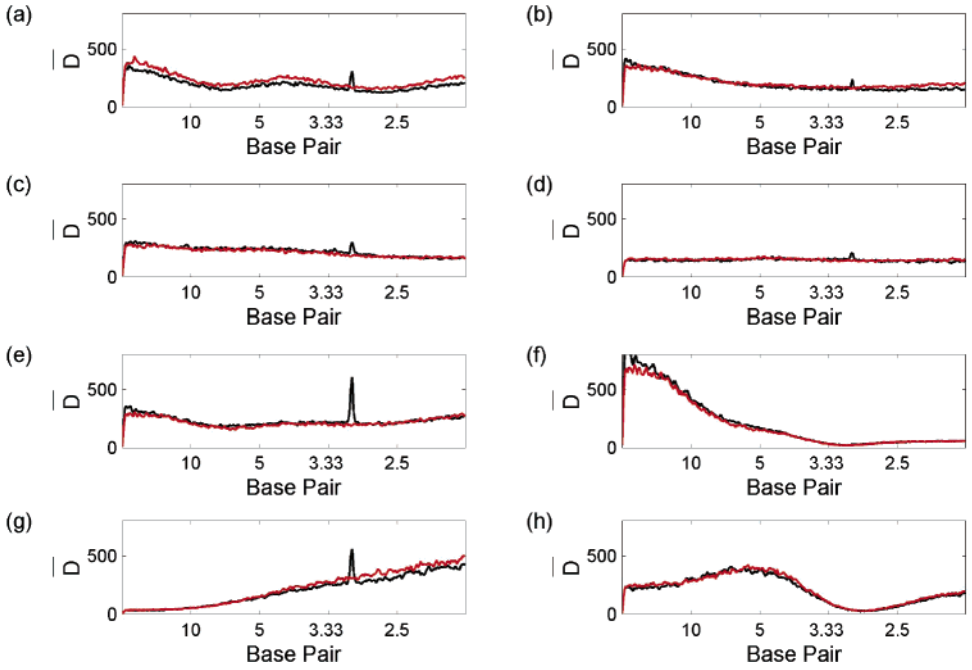

Figure 3. Structural power spectra of kCNG. $\bar{D}$ is the mean power spectra. Red line, randomized kCNG; black line, genomic kCNG. (a) Decreasing twist flexibility, (b) increasing twist flexibility, (c) decreasing roll flexibility, (d) increasing roll flexibility, (e) twist, (f) three-step twist, (g) roll, (h) three-step roll.

of analysis which might be carried out in detail for sets of sequences which exhibit some periodicity, rather than being a definitive algorithm. We then aligned each nUCE in turn to the ideal three-step roll trace by finding its best matching segment of length L. The approach used for the alignment is to calculate the energy required to distort the DNA away from its minimum energy three-step roll pattern into the ideal 6.2 bp periodic three-step roll. The segment with the lowest bending energy is deemed to be the best match. This is done for a nUCE of length N as follows: (1) Construct a vector of three-step roll values, one for each of N-7 overlapping octamers. (2) For each of the $N-L$ segments of length L, calculate the energy required for each component octamer to change its three-step roll value to match the appropriate ideal three-step roll at its central step using its three-step roll flexibility parameters. Each octamer has both a decreasing three-step flexibility parameter k_1 and an increasing three-step flexibility parameter k_2. To move from a minimum energy three-step roll value r_0 to a smaller ideal value r_1, the energy required is $k_1(r_0 - r_1)^2$, whereas to move to a

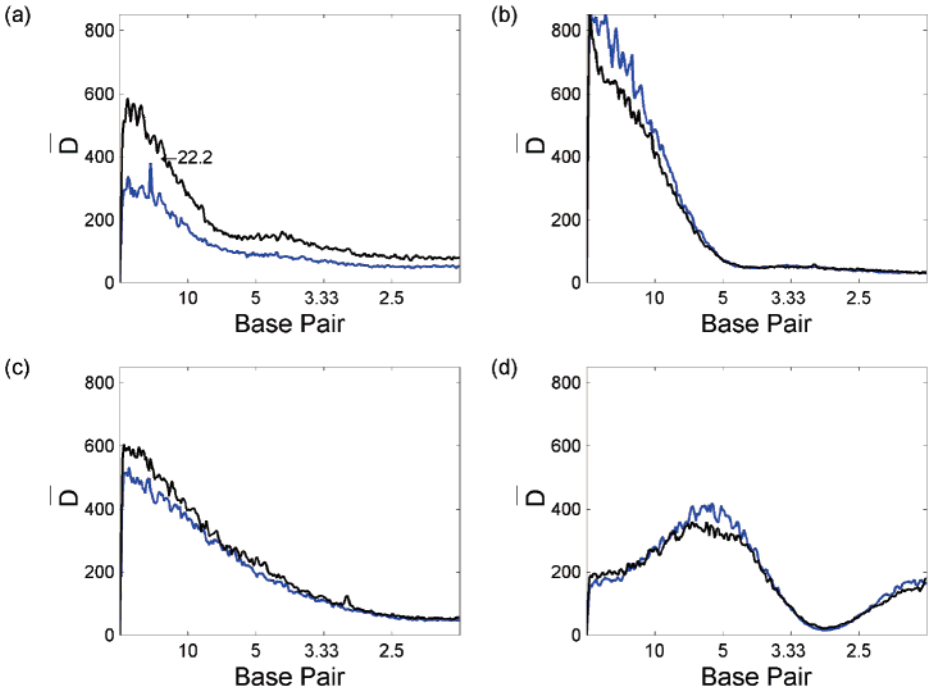

Figure 4. Structural power spectra of kCNG and uCNG. $\bar{D}$ is the mean power spectra; black line, kCNG; blue line, uCNG. (a) Decreasing three-step twist flexibility, (b) minor groove width, (c) decreasing three-step roll flexibility, (d) three-step roll.

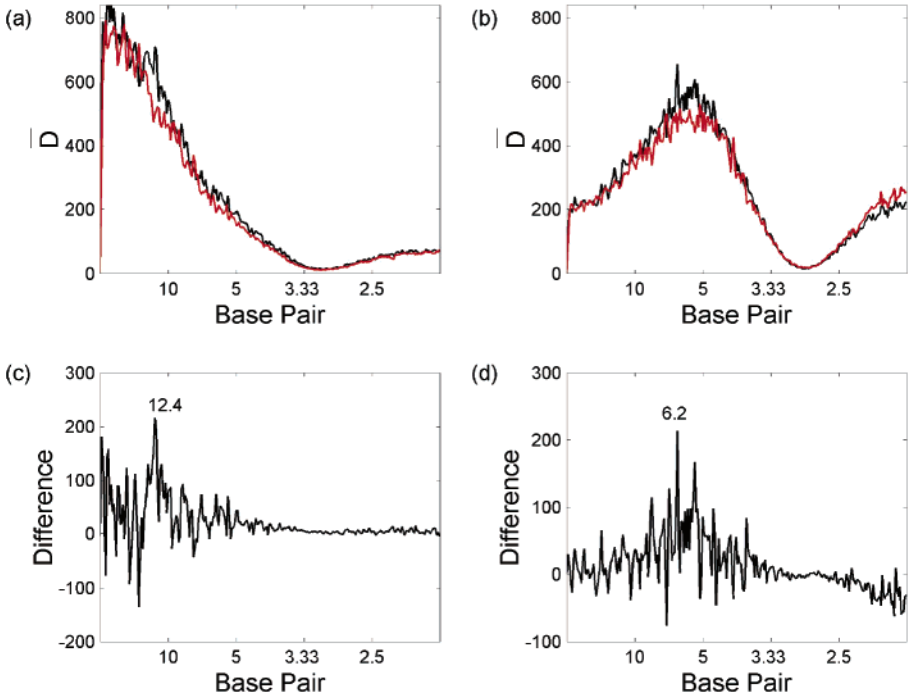

Figure 5. Three-step power spectra for nUCE. $\bar{D}$ is the mean power spectra. Red line, randomized nUCE; black line, genomic nUCE. (a) Three-step twist, (b) three-step roll, (c) difference spectrum for three-step twist, (d) difference spectrum for three-step roll.

larger ideal value r_2, the energy required is $k_2(r_0 - r_2)^2$. See Gardiner et al.[15] for further details. The sum of the energies gives a total strain energy for the segment. (3). The segment with the lowest energy requirement best matches the ideal three-step roll.

An ideal length $L = 60$ was chosen since we do not expect the periodic element to be very long, or necessarily of the same length in each nUCE. A length of 60 will be long enough to see if the periodicity can be located, and if the actual region of periodicity is longer, we will at least capture some of it. When an ideal length $L = 90$ was used, 6 of the 14 segments of length 90 entirely contained the 60mer previously identified, and a further 4 substantially overlapped the 60mer, demonstrating a degree of robustness in the method.

Figure 8a shows the three-step roll of the best-matching 60 bp segment of 1 of the 14 nUCEs (hg16_ct_Ultra_uc_448, whose power spectrum is plotted in Figure 7c), overlaid with the ideal three-step roll. The 6.2 bp periodicity that gives rise to the signal in the power spectrum in Figure 7c is clearly apparent over these 60 bases. In Figure 8b, we show the mean three-step roll of all 14 nUCEs, again overlaid with

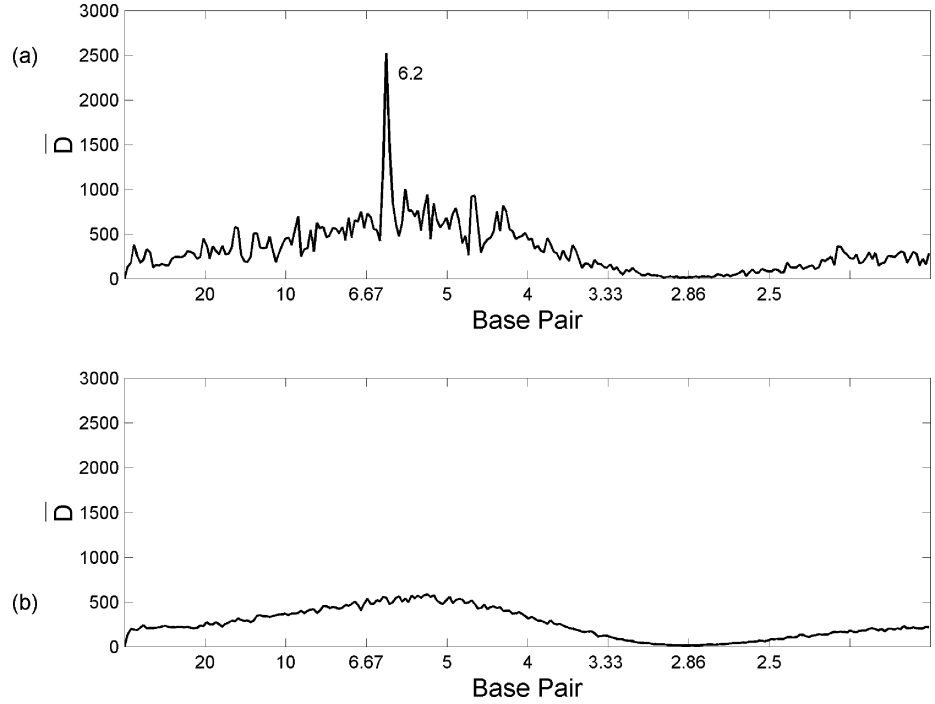

Figure 6. Comparison of three-step power spectra of special nUCE. (a) Mean spectra of 14 nUCE with three-step roll peak greater than mean + two standard deviations at 6.2 bp. (b) Mean spectra of 242 nUCE with three-step roll peak less than mean + two standard deviations at 6.2 bp.

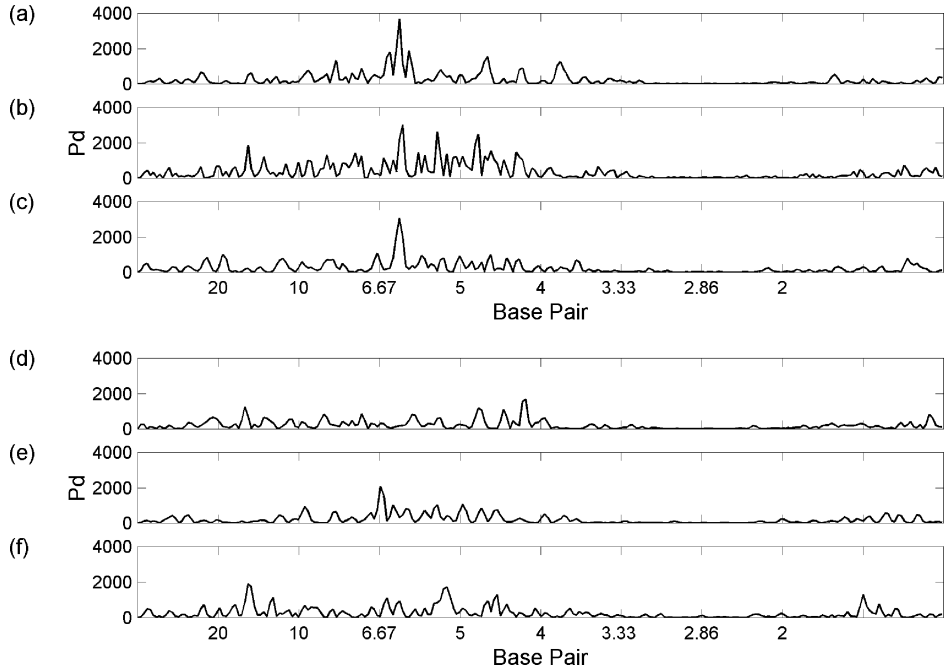

Figure 7. Power spectra of nUCE which do and do not exhibit three-step roll periodicity. Pd is the power spectral density; (a−c) spectra for three nUCE with 6.2 bp periodicity (identifiers hg16_ct_Ultra_uc_322, hg16_ct_Ultra_uc_445, hg16_ct_Ultra_uc_448); (d−f) spectra for three nUCE with no 6.2 bp periodicity (identifiers hg16_ct_Ultra_uc_5, hg16_ct_Ultra_uc_9, hg16_ct_Ultra_uc_15).

the ideal three-step roll. Averaging smooths the signal, and the fit between the real and ideal three-step roll is then very striking.

Of the 14 nUCEs, 9 appeared by eye to have common features. Figure 9 shows the three-step roll profiles of these nine. Rather than actual values, their *z* scores are plotted. The *z* score, $(x - \bar{x})/\sigma$, measures the significance of the three-step roll value x, where $\bar{x}$ is the mean and σ the standard deviation of the three-step roll over the octamer population. The color of the dots indicates the likelihood of finding a value of that size by chance. Blue dots represent values which are between one and two standard deviations from the mean, and yellow dots represent values which are more than two standard deviations from the mean.

Although the alignment method is extremely crude, it does show these nine nUCEs have a high peak (colored blue or yellow) in three-step roll at positions 49 or 50 and all bar one have another at position 24 or 25, four complete periods earlier. Eight also have a low value at position 45 or 46. The conservation of these extreme values suggests that the

Table 3. nUCE with 6.2 bp Three-Step Roll Periodicity

nUCE identifier	60 bp subsequence with best 6.2 bp three-step roll periodicity
hg16_ct_Ultra_uc_10	CTAAGTACTGTATTTTTTCACTCAAAAATTGTGAGTATAATGTAGTAATGATATCAAAAC
hg16_ct_Ultra_uc_39	CTTGTCTGTTTGAGGCCTACAGATGATGCTAAAGAGCCAATAATGACAGCAATTCAAAGC
hg16_ct_Ultra_uc_300	TTTTCCTGTGTTCCCATTATATTTGTTTGTAAAAGTGGAAATCAATTTTGAAGGTAATTT
hg16_ct_Ultra_uc_322	AATTTCTCCTTTTCATTTAATAGTAGACATTCATTAGGACCCACACACTGTAATAAATTA
hg16_ct_Ultra_uc_373	TTAAGATGCCATATCCCTTCTTGTAGCAATAAAGGGGTAAAGCAACATTAATATTAAATA
hg16_ct_Ultra_uc_445	TTTATCAAGAAATGGTTGTTTGCTGAGGTTATGTTTTGTCCATGAAATATGCATTTCATA
hg16_ct_Ultra_uc_448	AATTTTGTTCTGTGCCACAAATATAAAATTGTCCCTTGCAGAGAAATGTCAGCTTGCTAG
hg16_ct_Ultra_uc_60	TGCCCAGGGTTGTATTTGCATATGATATTCAGGGCATGATTTTTTTATTGTTCTTAATCA
hg16_ct_Ultra_uc_80	TAATTATAATGCTACAATAACATATTACGATGCCAGAACATATTACTGTACATTCTGTTA
hg16_ct_Ultra_uc_123	TAATGCTATTTCATTATCTGTTTGTATTTTATATTGACAGTGTCAGATAAACAGTGTGGA
hg16_ct_Ultra_uc_146	CATCTAAAACATGCCACCGTATGCAACACCACTTTATTATAAGTCACTGAGAGCAATATC
hg16_ct_Ultra_uc_206	ATTGGCGACCTTTCCTATTGTGTAGTGCTTTATGGTGTAATGGAAAGCGATCTTTACCAA
hg16_ct_Ultra_uc_227	AGCAGCATGGTTACATCTACACTAATAAAGAATTTTCAGCCTATAATCCAAAACAGAACC
hg16_ct_Ultra_uc_461	TTTGGCCCTCCACTAGACACTCTTGGGATTGTGCTTTTACAATAAATGTGTCTGTGATAG

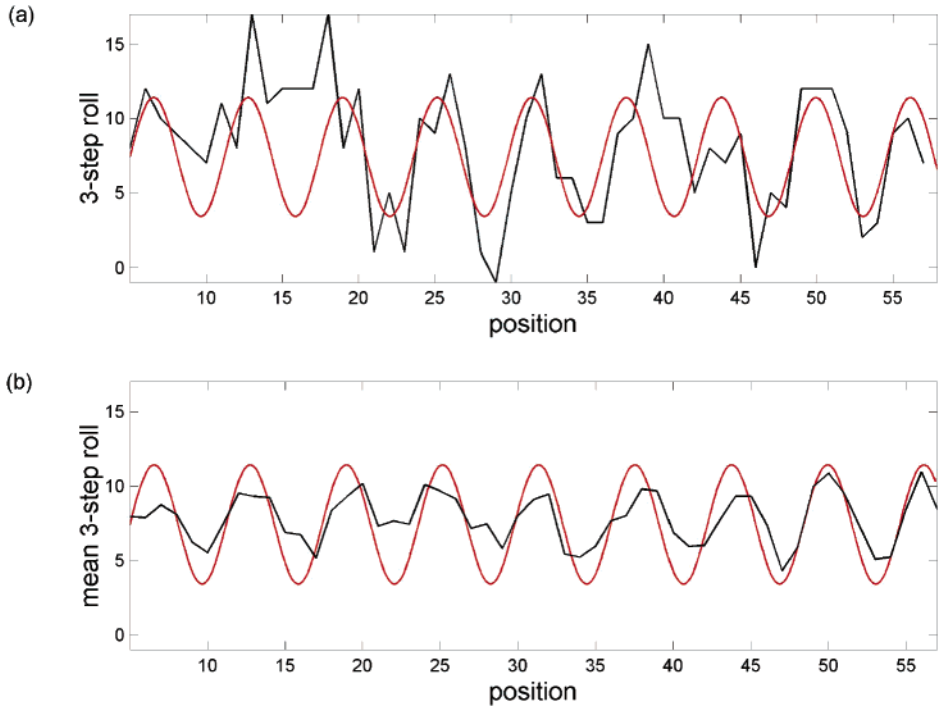

Figure 8. Section of nUCE hg16_ct_Ultra_uc_448 with 6.2 bp three-step roll periodicity. (a) Black line, nUCE (hg16_ct_Ultra_uc_448) three-step roll; red line, ideal three-step roll. (b) Black line, mean three-step roll for 14 nUCEs; red line, ideal three-step roll.

alignment is correct, at least for these nUCEs, and also indicates that they are likely to be of structural importance.

CONCLUSIONS

In this paper, we have described the use of Fourier transforms to search for patterns in the structural properties of highly conserved genomic elements of unknown function. Since neither the CNGs or the UCEs we considered possess any detectable sequence homology between paralogous elements in either species (human or mouse), such alignment-free techniques clearly offer the possibility of insight into the functional roles of these sequences.

Our initial aim was to find structural periodicities common to both the CNGs and UCEs, but although individual sequences may share common features, our techniques have not found any overall commonalities. One possible explanation for this is that the nUCEs tend to be found in introns or in regions known to be associated with transcription or regulation,[4] whereas the CNGs are largely randomly arranged in intergenic regions.[3] It is likely that the requirements of these different genomic regions give rise to different structural properties or that the observed structural properties reflect different functions. Fourier analysis of the structural properties of entire genomes and particular genomic regions is part of an ongoing investigation of DNA structure/function relationships, the results of which will be presented shortly.

We also analyzed the CNGs and UCEs separately, aiming to discover structural properties common to either the CNGs or the UCEs. The occurrence spectra, obtained by Fourier analysis of the sequence properties of the kCNG demonstrated a peak, as expected, at 3 bp, indicating the presence of codons. It is extremely encouraging that the Fourier analysis of the structural properties of the kCNG also revealed a peak at 3 bp. This demonstrates that the Fourier techniques explored here can reveal known features and also gives confidence that more unexpected peaks also correspond to real variation of the DNA structure.

Some uCNGs show periodicity in decreasing twist flexibility of ~22.2 bp, which is approximately twice the double-helical pitch of A-DNA.[22] This periodicity is not present in

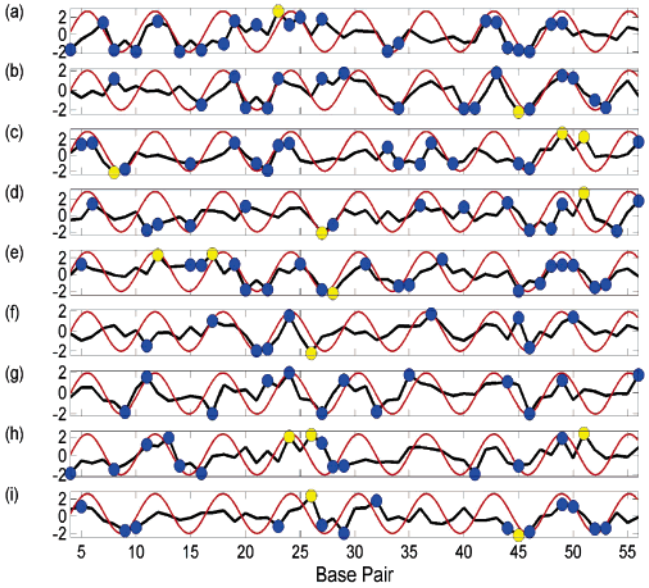

Figure 9. Three-step roll profiles. *Y* axis, *z* score, $(x - \bar{x})/\sigma$, where x = three-step roll; $\bar{x}$ is the mean and σ the standard deviation of the three-step roll over the octamer population. Points along each profile have been colored blue between one and two standard deviations from the mean and yellow for more than two standard deviations. Black line, *z* scores; red line, normalized ideal three-step roll. The sequences are (a) hg16_ct_Ultra_uc_10, (b) hg16_ct_Ultra_uc_300,(c)hg16_ct_Ultra_uc_322,(d)hg16_ct_Ultra_uc_445, (e)hg16_ct_Ultra_uc_448,(f)hg16_ct_Ultra_uc_80,(g)hg16_ct_Ultra_uc_123, (h) hg16_ct_Ultra_uc_206, (i) hg16_ct_Ultra_uc_227.

the sequence spectra, demonstrating that there is information present in the three-dimensional structure of double-helical DNA which is not apparent from the sequence.

Some nUCEs show a periodicity of 6.2 bp in three-step roll and also of 12.4 bp in three-step twist. This correlation means that there is likely to be some structural reason for this periodicity. However, the periodicity of 6.2 bp in three-step roll is out of phase with the 10.2 bp periodicity of the helical twist of DNA, so there are no dramatic consequences for the path adopted by the DNA; that is, it does not bend, in contrast to the nucleosome example discussed in the Introduction. The 6.2 bp periodicity of three-step roll is in phase with the helical repeat over ~30 bases, which is unlikely to cause significant bending within the DNA persistence length (150 bp). Thus, although we have identified a structural feature of these sequences, there are no readily apparent functional corollaries. It is possible that local variations in structure associated with protein binding over tens of bases give rise to the periodic signals detected here. For example, in our alignment in Figure 9, the region between bases 40 and 55 shows a pronounced local bend, with a relatively low three-step roll at about 46 bp followed by a high roll at about 49 bp.

This work demonstrates that it is possible to find structural patterns in DNA that are not observable as patterns in the nucleotide arrangements of the sequence. Although it is not yet possible to deduce a function from these patterns, the power spectrum approach presented here represents a useful method for identifying and classifying sequences.

ACKNOWLEDGMENT

We thank the BBSRC for support of this work and the Wolfson Foundation and the Royal Society for the provision of computing facilities. We thank Dr. Martin Whittle for useful discussions on the use of Fourier techniques for signal analysis.

REFERENCES AND NOTES

(1) Waterston, R. H.; Lindblad-Toh, K.; Birney, E.; Rogers, J.; Abril, J. F.; Agarwal, P.; Agarwala, R.; Ainscough, R.; Alexandersson, M.; An, P.; Antonarakis, S. E.; Attwood, J.; Baertsch, R.; Bailey, J.; Barlow, K.; Beck, S.; Berry, E.; Birren, B.; Bloom, T.; Bork, P.; Botcherby, M.; Bray, N.; Brent, M. R.; Brown, D. G.; Brown, S. D.; Bult, C.; Burton, J.; Butler, J.; Campbell, R. D.; Carninci, P.; Cawley, S.; Chiaromonte, F.; Chinwalla, A. T.; Church, D. M.; Clamp, M.; Clee, C.; Collins, F. S.; Cook, L. L.; Copley, R. R.; Coulson, A.; Couronne, O.; Cuff, J.; Curwen, V.; Cutts, T.; Daly, M.; David, R.; Davies, J.; Delehaunty, K. D.; Deri, J.; Dermitzakis, E. T.; Dewey, C.; Dickens, N. J.; Diekhans, M.; Dodge, S.; Dubchak, I.; Dunn, D. M.; Eddy, S. R.; Elnitski, L.; Emes, R. D.; Eswara, P.; Eyras, E.; Felsenfeld, A.; Fewell, G. A.; Flicek, P.; Foley, K.; Frankel, W. N.; Fulton, L. A.; Fulton, R. S.; Furey, T. S.; Gage, D.; Gibbs, R. A.; Glusman, G.; Gnerre, S.; Goldman, N.; Goodstadt, L.; Grafham, D.; Graves, T. A.; Green, E. D.; Gregory, S.; Guigo, R.; Guyer, M.; Hardison, R. C.; Haussler, D.; Hayashizaki, Y.; Hillier, L. W.; Hinrichs, A.; Hlavina, W.; Holzer, T.; Hsu, F.; Hua, A.; Hubbard, T.; Hunt, A.; Jackson, I.; Jaffe, D. B.; Johnson, L. S.; Jones, M.; Jones, T. A.; Joy, A.; Kamal, M.; Karlsson, E. K.; Karolchik, D.; Kasprzyk, A.; Kawai, J.; Keibler, E.; Kells, C.; Kent, W. J.; Kirby, A.; Kolbe, D. L.; Korf, I.; Kucherlapati, R. S.; Kulbokas, E. J.; Kulp, D.; Landers, T.; Leger, J. P.; Leonard, S.; Letunic, I.; Levine, R.; Li, J.; Li, M.; Lloyd, C.; Lucas, S.; Ma, B.; Maglott, D. R.; Mardis, E. R.; Matthews, L.; Mauceli, E.; Mayer, J. H.; McCarthy, M.; McCombie, W. R.; McLaren, S.; McLay, K.; McPherson, J. D.; Meldrim, J.; Meredith, B.; Mesirov, J. P.; Miller, W.; Miner, T. L.; Mongin, E.; Montgomery, K. T.; Morgan, M.; Mott, R.; Mullikin, J. C.; Muzny, D. M.; Nash, W. E.; Nelson, J. O.; Nhan, M. N.; Nicol, R.; Ning, Z.; Nusbaum, C.; O'Connor, M. J.; Okazaki, Y.; Oliver, K.; Overton-Larty, E.; Pachter, L.; Parra, G.; Pepin, K. H.; Peterson, J.; Pevzner, P.; Plumb, R.; Pohl, C. S.; Poliakov, A.; Ponce, T. C.; Ponting, C. P.; Potter, S.; Quail, M.; Reymond, A.; Roe, B. A.; Roskin, K. M.; Rubin, E. M.; Rust, A. G.; Santos, R.; Sapojnikov, V.; Schultz, B.; Schultz, J.; Schwartz, M. S.; Schwartz, S.; Scott, C.; Seaman, S.; Searle, S.; Sharpe, T.; Sheridan, A.; Shownkeen, R.; Sims, S.; Singer, J. B.; Slater, G.; Smit, A.; Smith, D. R.; Spencer, B.; Stabenau, A.; Stange-Thomann, N.; Sugnet, C.; Suyama, M.; Tesler, G.; Thompson, J.; Torrents, D.; Trevaskis, E.; Tromp, J.; Ucla, C.; Ureta-Vidal, A.; Vinson, J. P.; von Niederhausern, A. C.; Wade, C. M.; Wall, M.; Weber, R. J.; Weiss, R. B.; Wendl, M. C.; West, A. P.; Wetterstrand, K.; Wheeler, R.; Whelan, S.; Wierzbowski, J.; Willey, D.; Williams, S.; Wilson, R. K.; Winter, E.; Worley, K. C.; Wyman, D.; Yang, S.; Yang, S. P.; Zdobnov, E. M.; Zody, M. C.; Lander, E. S. Initial sequencing and comparative analysis of the mouse genome. *Nature* **2002**, *420*, 520−562.

(2) Dermitzakis, E. T.; Reymond, A.; Lyle, R.; Scamuffa, N.; Ucla, C.; Deutsch, S.; Stevenson, B. J.; Flegel, V.; Bucher, P.; Jongeneel, C. V.; Antonarakis, S. E. Numerous potentially functional but nongenic conserved sequences on human chromosome 21. *Nature* **2002**, *420*, 578−582.

(3) Dermitzakis, E. T.; Kirkness, E.; Schwarz, S.; Birney, E.; Reymond, A.; Antonarakis, S. E. Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **2004**, *14*, 852−859.

(4) Bejerano, G.; Pheasant, M.; Makunin, I.; Stephen, S.; Kent, W. J.; Mattick, J. S.; Haussler, D. Ultraconserved elements in the human genome. *Science* **2004**, *304*, 1321−1325.

(5) Dermitzakis, E. T.; Reymond, A.; Antonarakis, S. E. Conserved nongenic sequences—an unexpected feature of mammalian genomes. *Nat. Rev. Genet.* **2005**, *6*, 151−157.

(6) Antonarakis, S. E.; Lyle, R.; Dermitzakis, E. T.; Reymond, A.; Deutsch, S. Chromosome 21 and Down syndrome: From genomics to pathophysiology. *Nat. Rev. Genet.* **2004**, *5*, 725−738.

(7) Johnston, M.; Stormo, G. D. Heirlooms in the attic. *Science* **2003**, *302*, 997−998.

(8) Woolfe, A.; Goodson, M.; Goode, D. K.; Snell, P.; McEwen, G. K.; Vavouri, T.; Smith, S. F.; North, P.; Callaway, H.; Kelly, K.; Walter, K.; Abnizova, I.; Gilks, W.; Edwards, Y. J. K.; Cooke, J. E.; Elgar, G. Highly conserved noncoding sequences are associated with vertebrate development. *PLoS Biol.* **2005**, *3*, article number e7.

(9) Hunter, C. A. Sequence-Dependent Dna-Structure—the Role of Base Stacking Interactions. *J. Mol. Biol.* **1993**, *230*, 1025−1054.

(10) Hunter, C. A.; Lu, X. J. Construction of double-helical DNA structures based on dinucleotide building blocks. *J. Biomol. Struct. Dyn.* **1997**, *14*, 747−756.

(11) Hunter, C. A.; Lu, X. J. DNA base-stacking interactions: A comparison of theoretical calculations with oligonucleotide X-ray crystal structures. *J. Mol. Biol.* **1997**, *265*, 603−619.
(12) Packer, M. J.; Dauncey, M. P.; Hunter, C. A. Sequence-dependent DNA structure: Dinucleotide conformational maps. *J. Mol. Biol.* **2000**, *295*, 71−83.
(13) Packer, M. J.; Dauncey, M. P.; Hunter, C. A. Sequence-dependent DNA structure: Tetranucleotide conformational maps. *J. Mol. Biol.* **2000**, *295*, 85−103.
(14) Packer, M. J.; Hunter, C. A. Sequence-structure relationships in DNA oligomers: A computational approach. *J. Am. Chem. Soc.* **2001**, *123*, 7399−7406.
(15) Gardiner, E. J.; Hunter, C. A.; Packer, M. J.; Palmer, D. S.; Willett, P. Sequence-dependent DNA structure: A database of octamer structural parameters. *J. Mol. Biol.* **2003**, *332*, 1025−1035.
(16) Gardiner, E. J.; Hunter, C. A.; Lu, X. J.; Willett, P. A Structural Similarity Analysis of Double-Helical DNA. *J. Mol. Biol.* **2004**, *343*, 879−889.
(17) Luger, K.; Mader, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 angstrom resolution. *Nature* **1997**, *389*, 251−260.
(18) Fickett, J. W.; Tung, C. S. Assessment of Protein Coding Measures. *Nucleic Acids Res.* **1992**, *20*, 6441−6450.
(19) Tiwari, S.; Ramachandran, S.; Bhattacharya, A.; Bhattacharya, S.; Ramaswamy, R. Prediction of probable genes by Fourier analysis of genomic sequences. *Comput. Appl. Biosci.* **1997**, *13*, 263−270.
(20) Sharma, D.; Issac, B.; Raghava, G. P. S.; Ramaswamy, R. Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics* **2004**, *20*, 1405−1412.
(21) Silverman, B. D.; Linsker, R. A Measure of DNA Periodicity. *J. Theor. Biol.* **1986**, *118*, 295−300.
(22) Calladine, C. R.; Drew, H. R. *Understandng DNA*, 2nd ed.; Academic Press: Cambridge, U. K., 1997.

CI050384I