A gradient measure to characterize the interface between non-coding and coding regions in genomic DNA sequences

Prashant Goswami* and Sowmya Raghavan

CSIR Centre for Mathematical Modelling and Computer Simulation, Bangalore 560 037, India

Coding and non-coding sequences in DNA have long been characterized on the basis of their differences in base composition and degree of nucleotide variability. Here, we describe a position-dependent dinucleotide density function (NDF) to understand the characteristics of a typical interface between non-coding and coding regions in genomes. This function is observed to exhibit a sharp gradient uniquely at this interface. This gradient appears to be not only statistically robust, but also phylogenetically conserved. The absence of a simple correlation between the dinucleotide counts (or frequencies) and this distribution function suggests that the gradient is a characteristic of the interface and cannot be deduced from the overall composition of the sequence alone. Furthermore, a randomly generated sequence of the same dinucleotide composition as a given genome also does not show the appearance of sharp gradients, implying that the NDF signal is a unique property of the coding and non-coding interface. In the set of representative organisms studied in this work, the gradients of functions corresponding to AC, TA and CT appear to be most prominent and conserved. Genome-specific signals are also present for some organisms. Hence we propose the existence of a position-dependent property of some dinucleotides that shows a marked change at the start of protein-coding regions, which is presumably interpreted by the read-out machinery of the cell.

CODING and non-coding sequences in DNA have long been characterized on the basis of their differences in base composition and degree of nucleotide variability. Several attempts have been made to identify characteristics that distinguish coding from non-coding regions, especially since the availability of genome sequence data. These include the study of various nucleotide contents *vis-á-vis* codon usage and base composition[1,2]. Analyses of statistical measures of sequence data and the identification of signals related to gene expression have also been carried out[3]. More recent studies have also analysed the differences in standard deviations of the nucleotide distributions between coding and non-coding regions[4]. Although characteristics derived from these quantities are most widely exploited in the present-day gene-finding algorithms, none of these methods can be said to be self-sufficient[5]. In addition, there are no explicit studies that seek to understand whether the interface between coding and non-coding regions has any special features.

A primary source of difficulty, of course, is that there appears to be no pervading uniformity in the basic gene structure. One avenue for searching a solution to this problem is to explore and identify signals that characterize coding regions. For example, a method for the accurate identification of splice sites would enable us to delineate the boundaries of exons and introns. It would appear that hybrid techniques that combine statistical measures and framework models of gene structures would be more effective in recognizing coding sequences. While this may be true in terms of the discriminatory function of the technique, a different approach will be necessary to explore genes outside the known classes. This is because most of the methods such as those based on Hidden Markov models are severely biased by known classes of genes, since they attempt to determine, in a probabilistic manner, whether a given sequence belongs to a known class of genes[6].

It is possible, on the other hand, that gene sequences are distinguished by certain intrinsic properties that could be local, global or both. In particular, local properties are likely to play a dominant role, since a genomic sequence is distinguished by a number of special regions, including start and stop sites. Across these specialized regions, various physicochemical properties are likely to vary sharply. Assuming that these (unidentified) properties are reflected in the distribution of nucleotides in the sequence, an appropriately formulated nucleotide distribution function should reveal these specialized locations through the occurrence of sharp gradients at certain locations. Furthermore, if this is a property intrinsic to genes, then such gradients could serve as signals to distinguish a 'gene' sequence from a 'non-gene' sequence, even if no similar genes are available in the database.

The aim of the present work is to explore the existence of such a measure that characterizes the interface between non-coding and coding regions accurately. The emphasis of the work is thus on the transition from non-coding to coding region (henceforth referred to as the interface), rather than on the global characteristics of the two regions. We observed that a position-dependent mononucleotide distribution function is not enough to achieve this distinction, but that of a dinucleotide does so with sufficient clarity. We also observed that the distribution function for some dinucleotides shows a gradient that is independent of the genome being analysed and hence appears to be a universal characteristic of the interface. In addition, there appear to be some genome-specific signals that occur in the distribution function of

*For correspondence. (e-mail: goswami@cmmacs.ernet.in)

other dinucleotides that can be rationalized on the basis of the genomic base composition.

If we assume that the coding region is distinguished from the non-coding region by some local characteristics, then a potential candidate for such a characteristic is a gradient function. In other words, the transition from the non-coding to coding regions could be signalled by a sharp change in some quantity along the sequence; the force defined by the gradient then determines the response across the zone of transition. Although the nature of this quantity cannot be guessed *a priori*, it is clear that it must be a function of nucleotide density. We thus define a local nucleotide density function (NDF) as

$$\rho_k(n) = 1/n \sum_{i=1}^{n} \eta_k(i),$$

where i is the position in the DNA sequence measured from some reference point (origin) up to the location n and k is the nucleotide type. The nucleotide type includes not only the mononucleotides but higher orders as well. The quantity $\eta_{jk}(i)$ is calculated as follows:

$\eta_k(i) = 1$, if the nucleotide k is present at i and 0 otherwise.

The gradient of NDF can now be defined as

$$\Delta\rho_k(n) = \rho_k(n) - \rho_k(n-1),\ n = 1, 2, \ldots N,$$

where N is the total number of locations in the sequence.

The forms of NDF and the gradient function explicitly for the case of the dinucleotide are as follows:

$$\rho_{jk}(n) = 1/n \sum_{i=1}^{n} \eta_{jk}(i),$$

$$\Delta\rho_{jk}(n) = \rho_{jk}(n) - \rho_{jk}(n-1).$$

As in the first equation, the suffix of η (jk in this case) refers to the nucleotide type, two indices in the present case to indicate dinucleotide and i measures the position along the sequences. Thus $\rho_{jk}(n)$ represents cumulative dinucleotide density at location n along the sequence. It is possible to construct such functions for higher orders of nucleotides as well. However, due to the fact that a strong signal was seen for the dinucleotide function itself, analysis using such higher-order functions was not carried out in the present study. To examine the statistical robustness of the signal, a number of sequences were considered for each organism. The average of gradient function was calculated as follows:

$$\langle \Delta\rho(n) \rangle = 1/M \sum_{l=1}^{M} \Delta\rho_l(n),$$

where M is the number of sequences that were considered for each organism. In this report, M values from 50 to 500 have been considered.

Since we were interested in the boundaries between non-coding and coding regions in genomes, we chose a set of gene sequences from the recently sequenced bacteria to test this hypothesis[7–12]. The eubacteria *E. coli*, *M. tuberculosis* and *B. burgdorferri* were chosen, as they represent moderate (53%), extremely high (65.6%) and extremely low G + C (27%) content, respectively. *M. jannaschii* and *M. thermoautotrophicum* were chosen among the archaea to represent extremes of composition in this class and *S. cerevisiae* was chosen among the eukaryotes. The complete set of well-characterized genes was taken from each of these organisms and the sequences 200 nt prior to and 200 nt after the start codon were extracted using custom PERL programs. (The most common start codon is ATG; however, TTG and GTG are also found. The term pre-ATG is used in the text here to denote the sequences before all three types of start codons.) For each of the sequences (except the individual chromosomes of yeast), sample sets of size M ranging from 50 to 500 were constructed from the set of known (i.e. well-characterized) genes. It was found that the results do not show any appreciable variations within this range. All subsequent discussions pertain to a sample size of 500, unless otherwise stated. The results of the analysis were insensitive to the five hundred genes which were analysed. ORFs designated as hypothetical (including conserved hypotheticals) were not considered in this analysis.

After computing the dinucleotide density function for each of the organisms, its gradient along the position from -200 to $+200$ positions of ATG was plotted. These results for various organisms are shown in Figures 1–6. Variation in the gradients of NDF as a function of sequence locations, is immediately apparent from these figures. What is striking however, is the conservation of some extrema, i.e. the ones corresponding to AC, TA and CT. In other words, all organisms clearly show a prominent transition for the NDFs of AC, TA and CT at the ATG codon. The local extremum value of the gradient of NDF around the ATG position thus appears to be a universal signal denoting the start of the coding region. Although not apparent from the figures, the sharp change in the gradient at the ATG position can either be positive or negative, as a closer inspection of the region around the ATG position reveals (not shown here). The strength of the individual signals varies between genomes, however, there was no simple relationship with the base compositions. In addition, the changes in the gradient are non-existent at sites

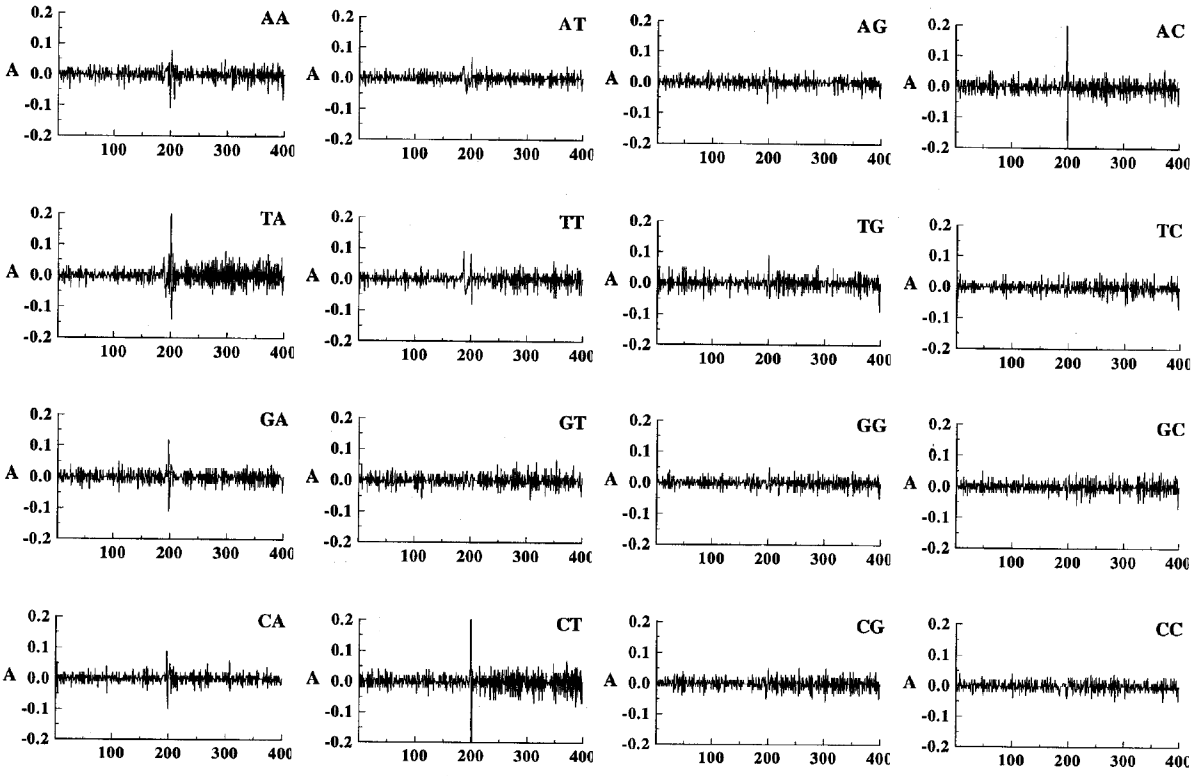

Figure 1. Plot of NDF gradient as a function of nucleotide position from 200 nucleotides prior to ATG up to 200 nucleotides post-ATG for the genome of *E. coli*.

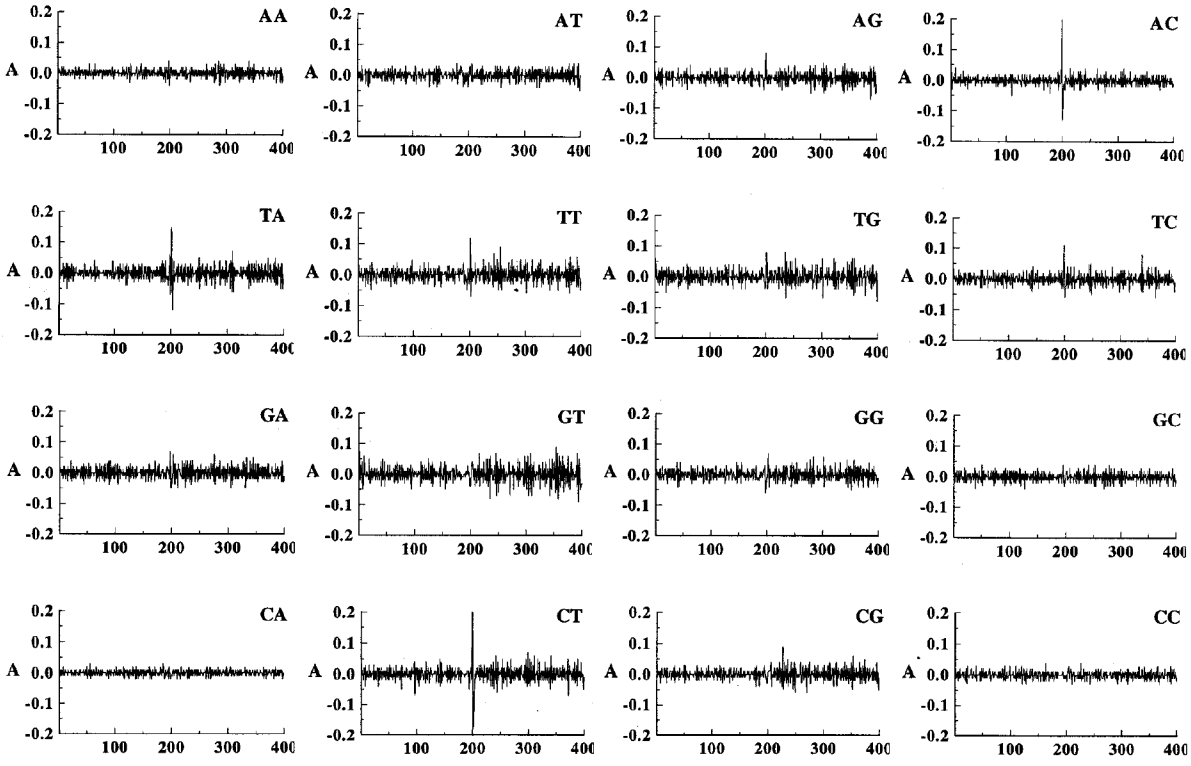

Figure 2. Plot of NDF gradient as a function of nucleotide position from 200 nucleotides prior to ATG up to 200 nucleotides post-ATG for the genome of *M. tuberculosis*.

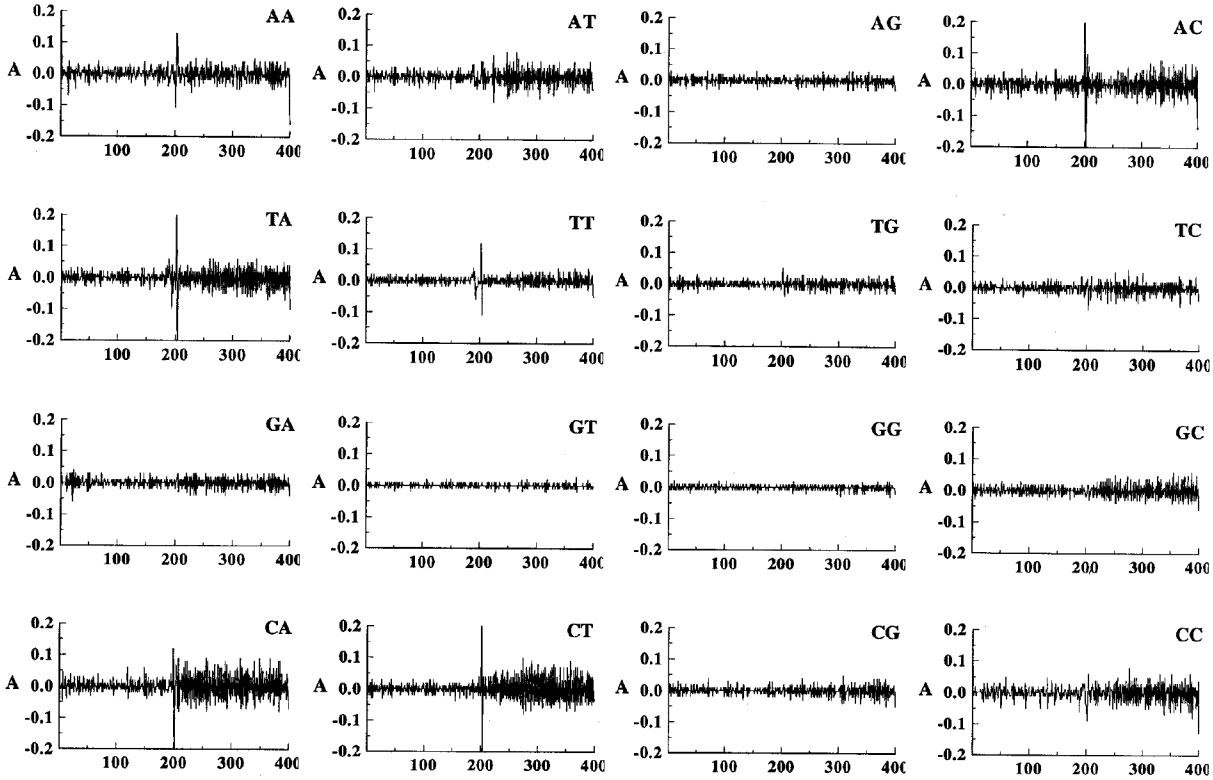

Figure 3. Plot of NDF gradient as a function of nucleotide position from 200 nucleotides prior to ATG up to 200 nucleotides post-ATG for the genome of *B. burgdorferi*.

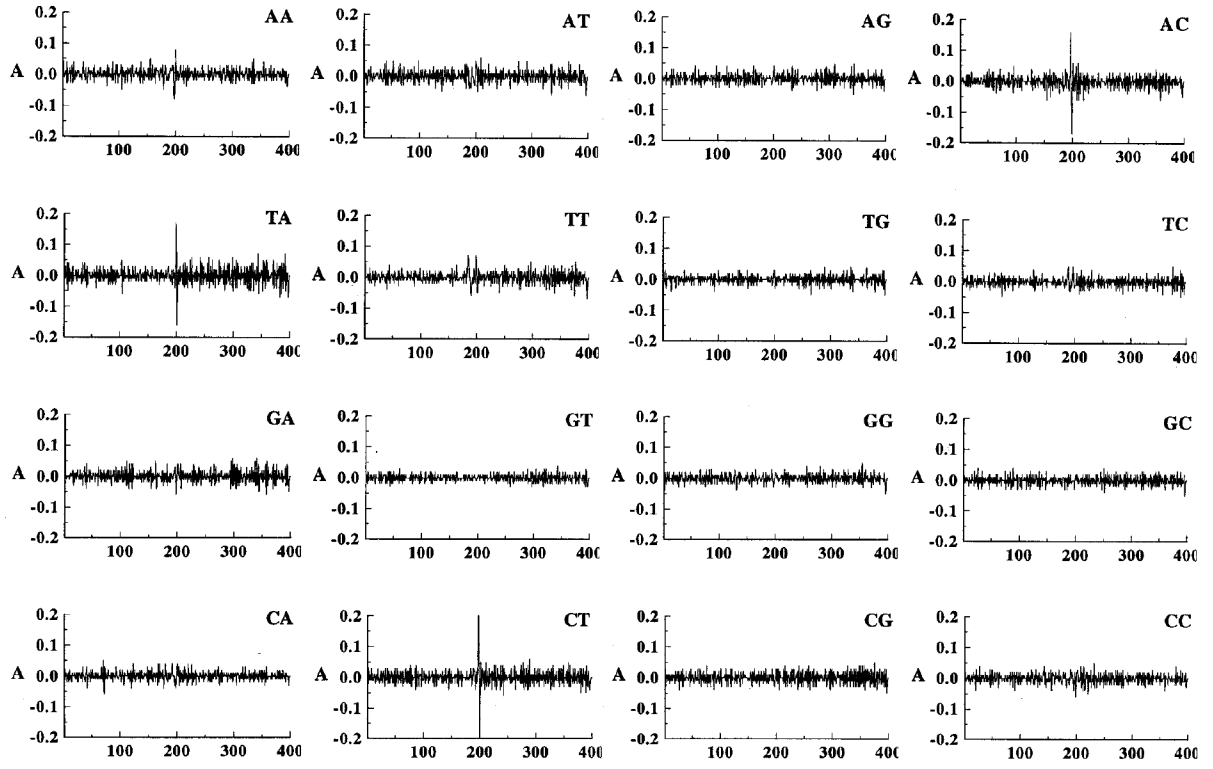

Figure 4. Plot of NDF gradient as a function of nucleotide position from 200 nucleotides prior to ATG up to 200 nucleotides post-ATG for the genome of *M. thermoautotrophicum*.

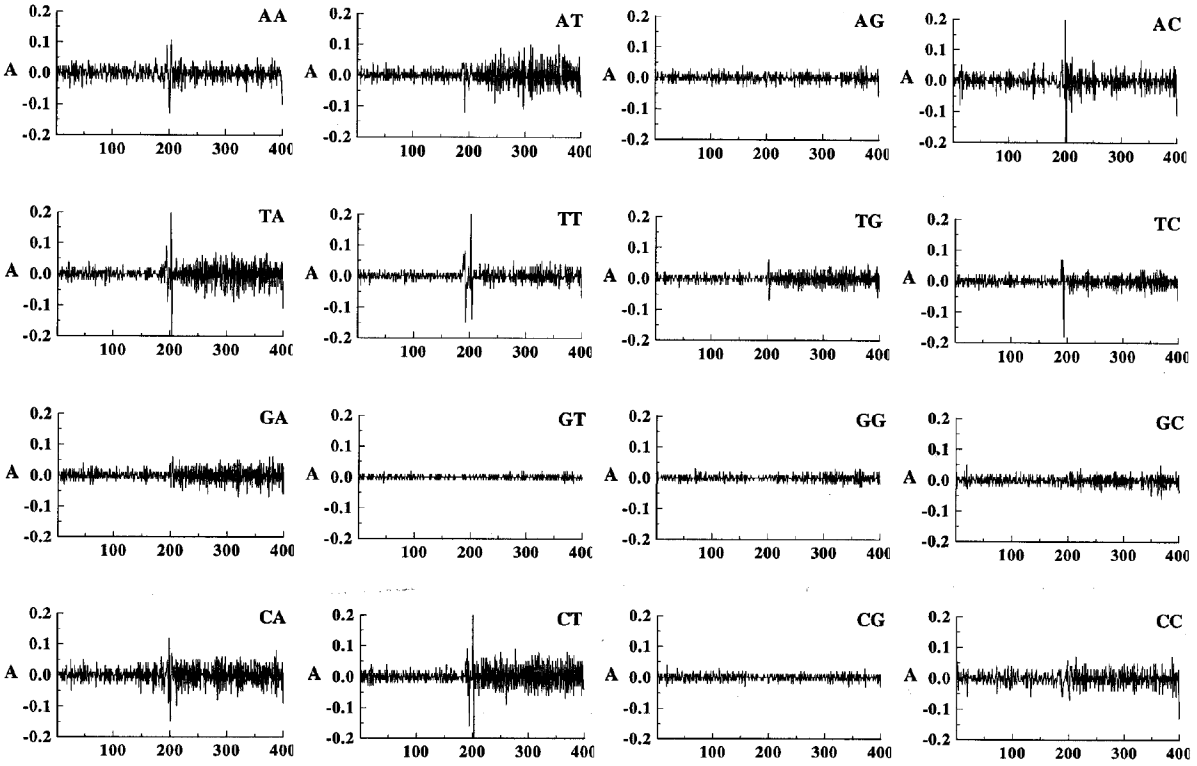

Figure 5. Plot of NDF gradient as a function of nucleotide position from 200 nucleotides prior to ATG up to 200 nucleotides post-ATG for the genome of *M. jannaschii*.

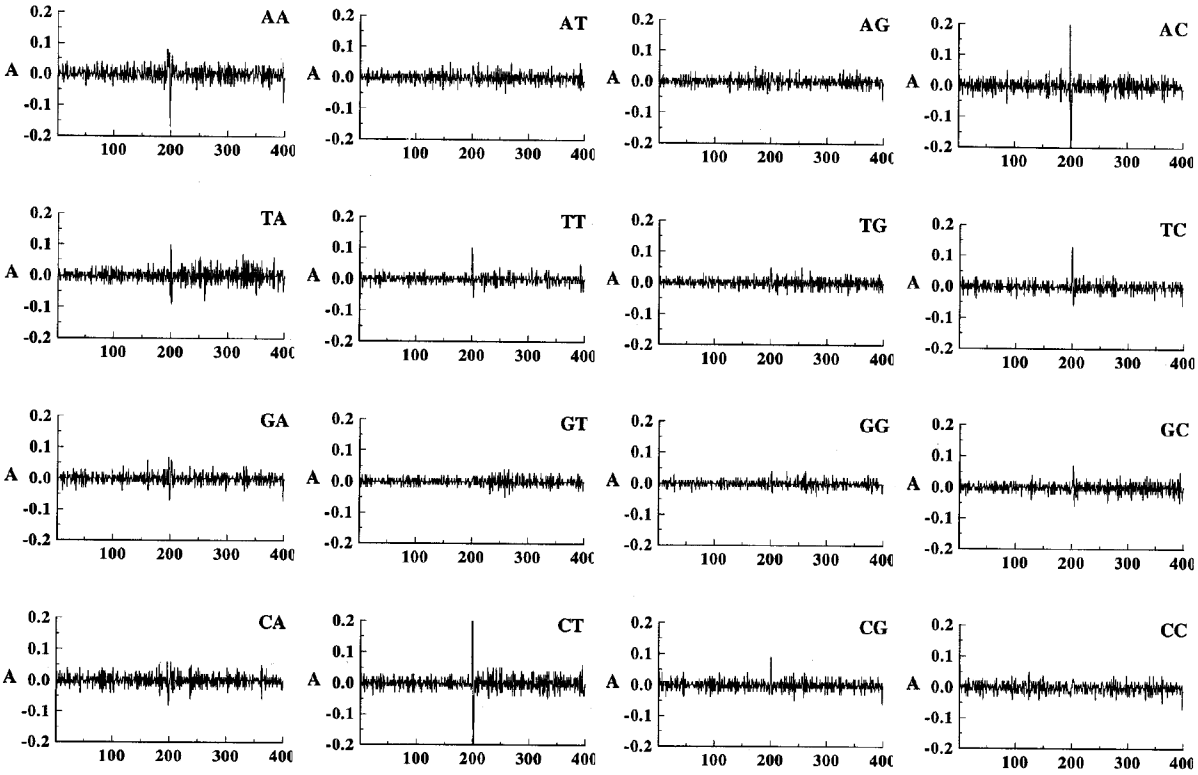

Figure 6. Plot of NDF gradient as a function of nucleotide position from 200 nucleotides prior to ATG up to 200 nucleotides post-ATG for the genome of *S. cerevisiae*.

far away from the position of ATG, i.e. well into the coding and the non-coding regions.

We rationalized that if the observed signal is unique to the interface, then it must not occur elsewhere in the sequence. We therefore performed two kinds of perturbation experiments. We generated random sequences 400 nt long, having the same dinucleotide content as a given genomic sequence, thus preserving the total content of dinucleotides. This led to a total loss of signal (data not shown), suggesting that the signal is unique to the interface. The loss of signal also indicates that the observed maxima are not due to dinucleotide composition alone. Next, we randomized the sequence before ATG and the sequence after ATG separately. In the former case, we generated 500 sequences containing the same dinucleotide composition as the non-coding region of a given genome and in the latter, of the same dinucleotide composition as the coding region. We observed that new maxima appeared in case of both the perturbations, with only the maximum in the TA being preserved. In any case, the loss of signals on this kind of perturbation further stresses the importance of the interface as a separate entity to be reckoned with.

In addition to the above perturbations, we also performed a simple count for the sixteen dinucleotides in the pre-ATG and post-ATG regions. We found that there was no simple relationship between the counts and the occurrence of the signal at the interface, i.e. the appearance of the signal did not correlate with a sudden increase or decrease in the number of dinucleotides of a given kind.

To ensure that the appearance of the signal is not due to the systematic occurrence of the Shine–Dalgarno sequence, we removed these signals from the pre-ATG sequences and repeated the analysis. The positions and strengths of the signals did not alter upon doing this, hence the signal cannot arise due to the systematic occurrence of this ribosome-binding motif.

Dinucleotide content has long been studied as a signature of genomic variation[13]. The abundance of purine : purine type of dinucleotide over purine : pyrimidine type has been rationalized on the basis of the fact that DNA structure tends to avoid clashes between opposite-strand nearest-neighbour purines[14]. Species-specific differences are therefore likely to be a complex function of genome packing and coding constitution[13]. In the analysis of NDFs too, we find that there are differences between genomes. We note that the AT-rich genomes such as *M. jannaschii* and *B. burgdorferi* have additional signals in AA and TT. The GA signal appears uniquely in *E. coli*. Even the simplest eukaryote, yeast, has chromosome-specific signals (data not shown). However, only a part of this can be rationalized on the basis of genome composition. For example, the signal in TA appears both in *M. tuberculosis*, a GC-rich genome and in *B. burgdorferi*, an AT-rich genome.

It is interesting to note that the variability in NDF is very small for most dinucleotides in the pre-ATG region. The enhanced variability in the post-ATG region is thus a characteristic of coding regions only and indicates a sparse use of that dinucleotide in the coding regions. An increase in variance is thus a good measure of coding region, as has also been reported earlier[4]. We also observed that the genomes analysed here differ in their correlation coefficients of pre- and post-ATG values of NDF. Thus the definition of NDF allows us to use a quantitative measure to distinguish between genomes.

The foregoing discussion pertains to the interface between non-coding and coding regions. The other canonical interface in genomes is the one between coding and non-coding regions, i.e. at the stop codon. To understand whether there is a sharp gradient in any of the NDFs at the stop codons, we plotted the gradient along the position from –200 to + 200 positions of TAG. These results for various organisms are shown in Figures 7–12. Once again, variations in the gradients of NDF as a function of sequence locations, are immediately apparent from these figures. What is striking however, is that the maxima that characterize the stop codon are in general different from those that characterize the start codon. The appearance of a new extremum corresponding to CA appears to be universal in that it is independent of genomic base composition. In addition, it is clear that the maxima corresponding to AC and CT are lost, collectively. In other words, either AC or CT maximum is present, but not both. This further stresses the importance of the gradient maxima as a unique property of the start interface. The TA maximum for the stop codon case appears only in somewhat AT-rich genomes such as the archaea. Thus it appears that the stop codon interface is also characterized by gradients in dinucleotide density functions.

To supplement our analysis of the gradient measure and compare our analysis with the statistical properties of genomic sequences, we calculated the standard deviations of the dinucleotide densities for the pre-ATG and post-ATG segments (prepared as described earlier). In general, all genomes exhibited greater variability in post-ATG regions than the pre-ATG regions. This is

Table 1. Ratio of variances of the dinucleotide density function

	Pre-ATG/post-ATG variance	
Organism	Minimum	Maximum
B. burgdorferi	0.05 (CT)	0.83 (AG)
E. coli	0.50 (CG)	2.89 (AC)
M. tuberculosis	0.40 (CG)	2.02 (AC)
M. thermoautotrophicum	0.60 (CG)	2.19 (CT)
M. jannaschii	0.11 (CT)	1.24 (CG)
S. cerevisiae	0.44 (TA)	3.69 (AC)

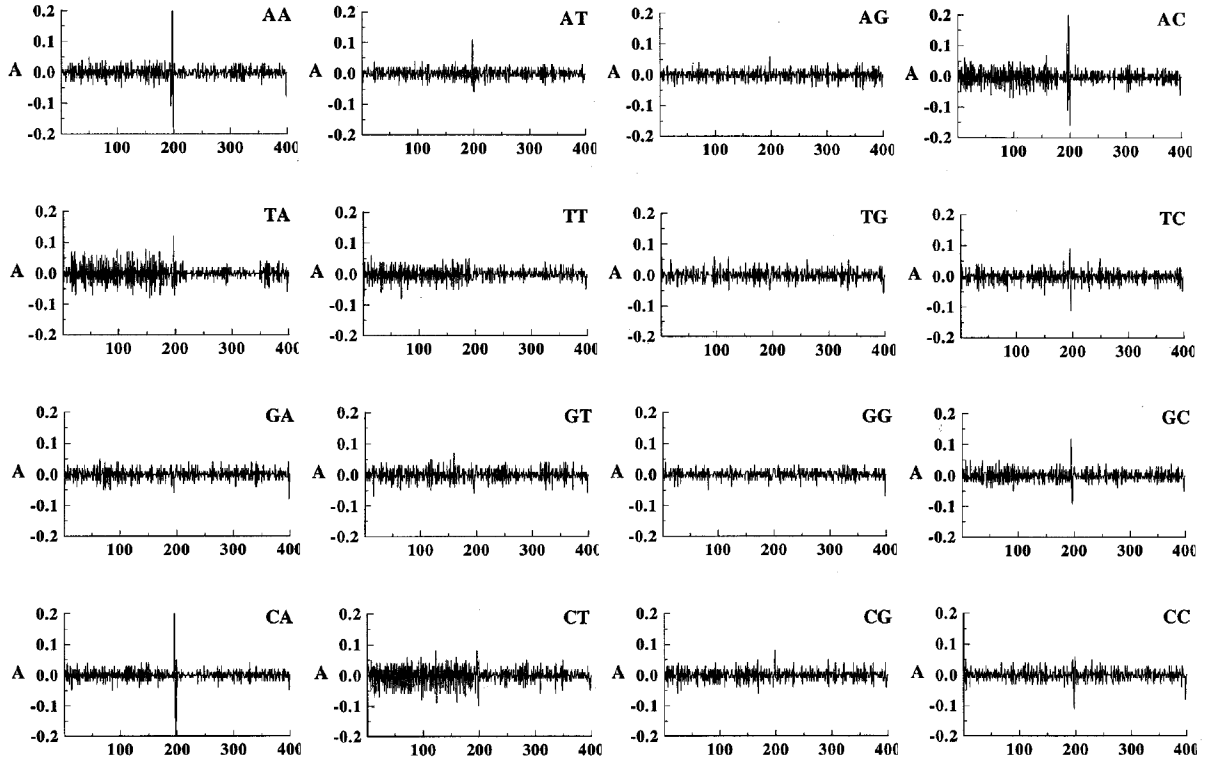

Figure 7. Plot of NDF gradient as a function of nucleotide position from 200 nucleotides prior to TAG up to 200 nucleotides post-TAG for the genome of *E. coli*.

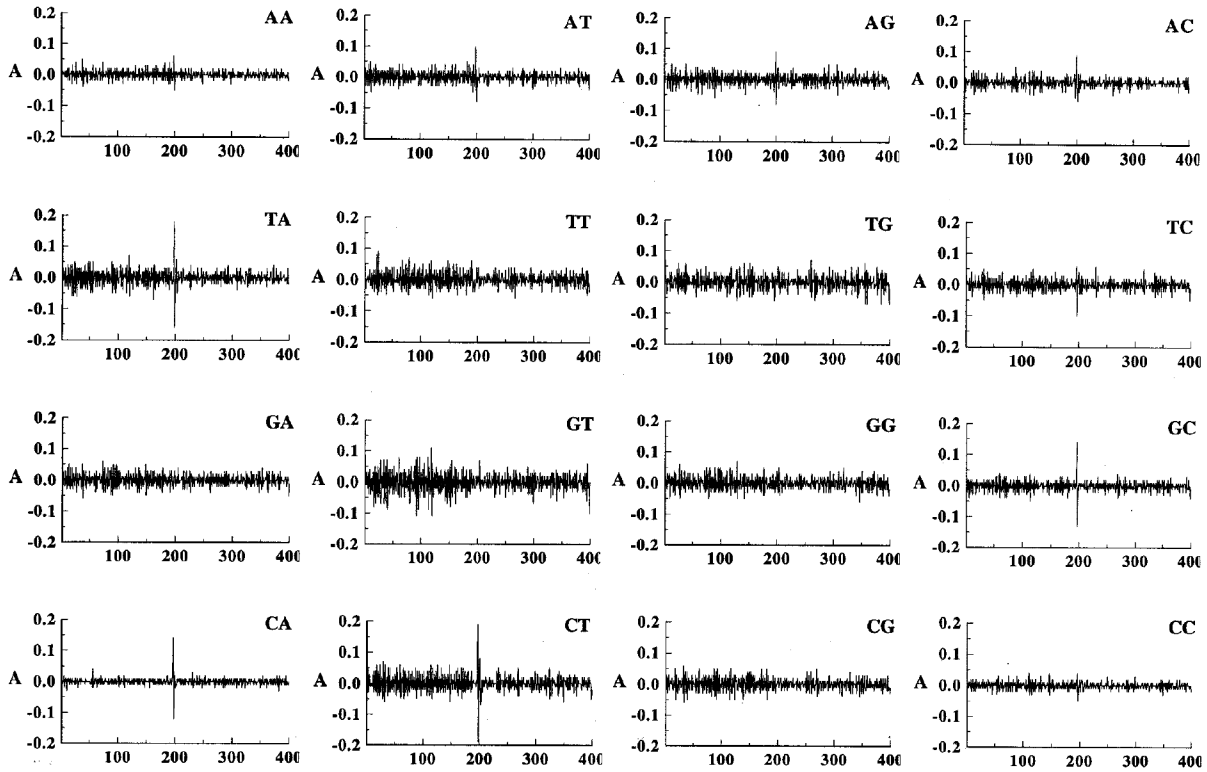

Figure 8. Plot of NDF gradient as a function of nucleotide position from 200 nucleotides prior to TAG up to 200 nucleotides post-TAG for the genome of *M. tuberculosis*.

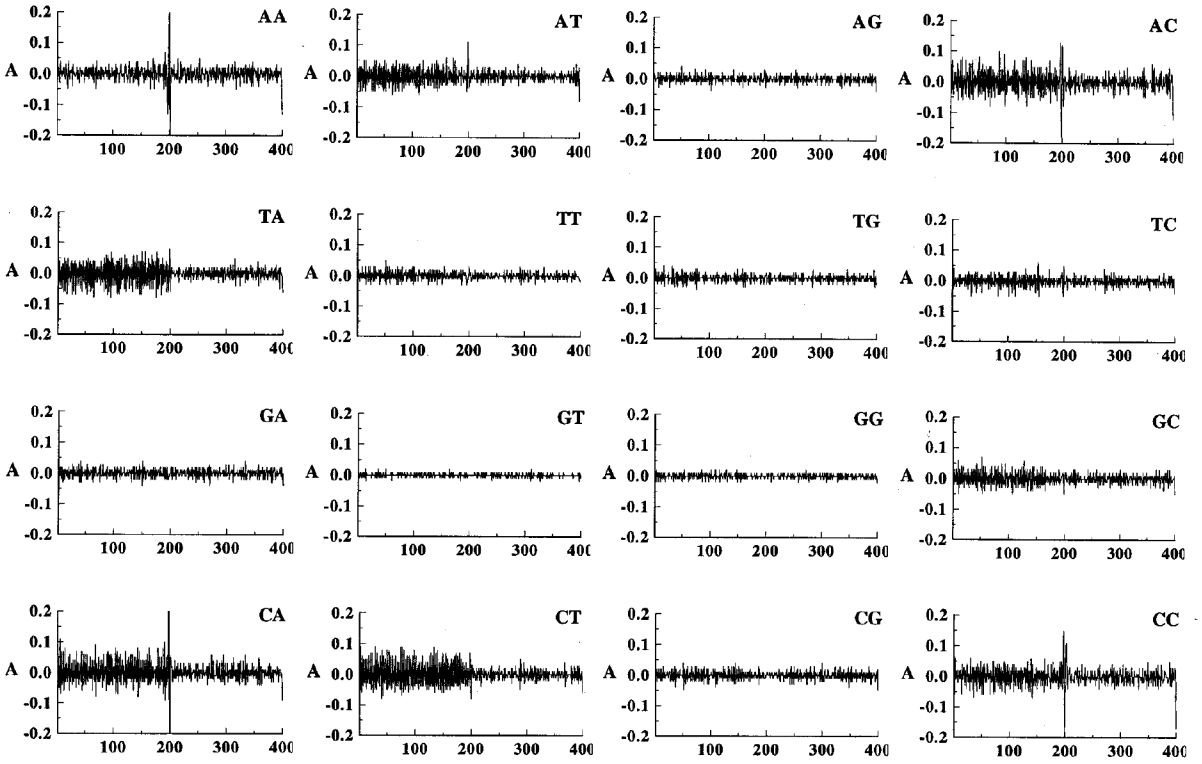

Figure 9. Plot of NDF gradient as a function of nucleotide position from 200 nucleotides prior to TAG up to 200 nucleotides post-TAG for the genome of *B. burgdorferi*.

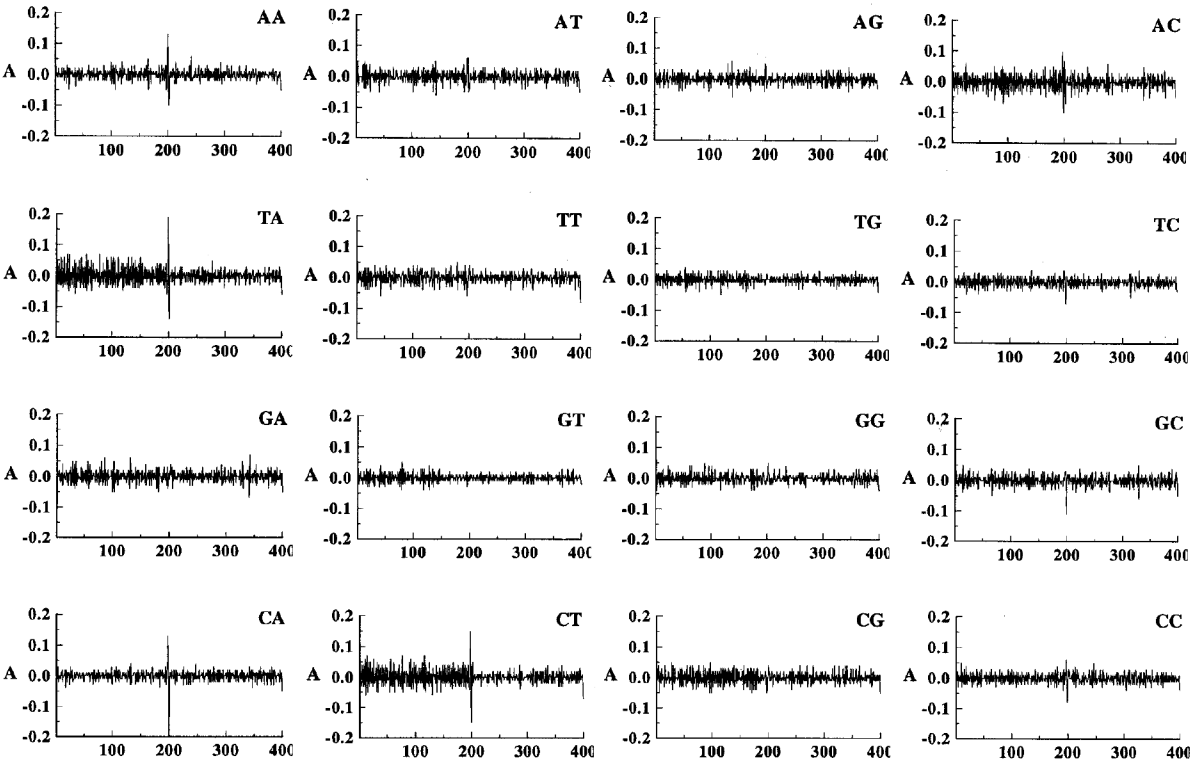

Figure 10. Plot of NDF gradient as a function of nucleotide position from 200 nucleotides prior to TAG up to 200 nucleotides post-TAG for the genome of *M. thermoautotrophicum*.

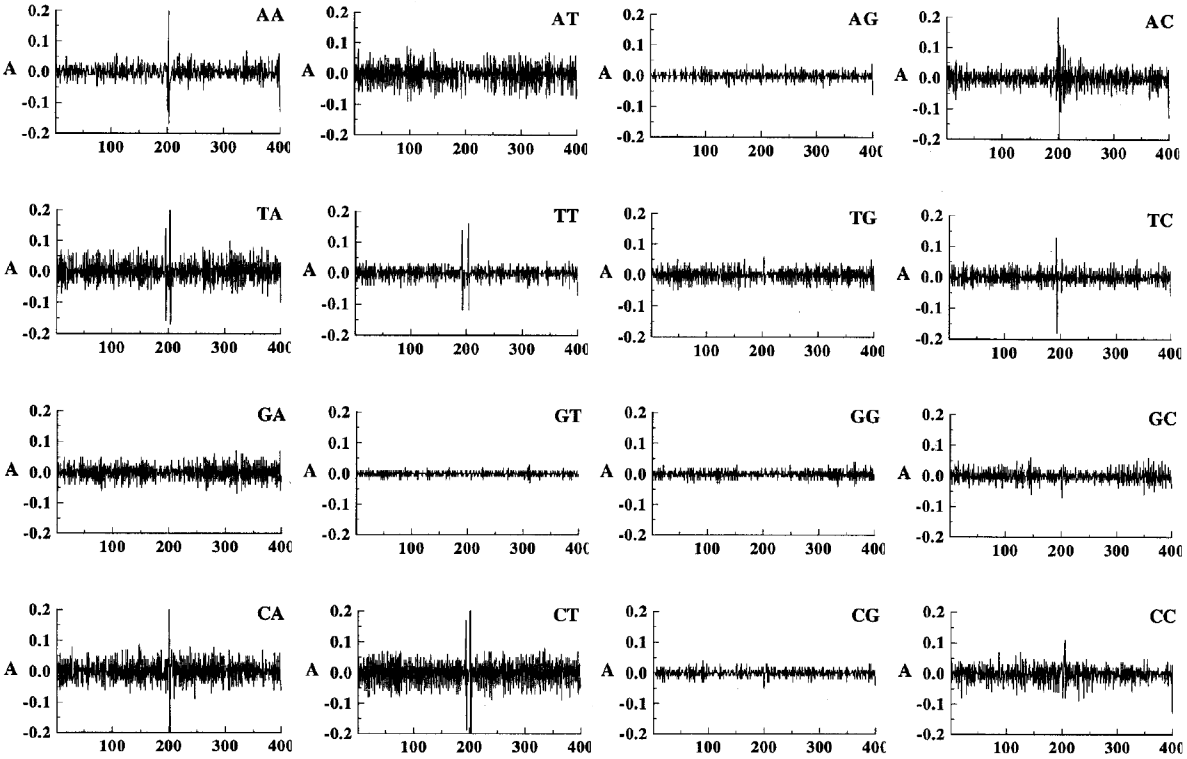

Figure 11. Plot of NDF gradient as a function of nucleotide position from 200 nucleotides prior to TAG up to 200 nucleotides post-TAG for the genome of *M. jannaschii*.

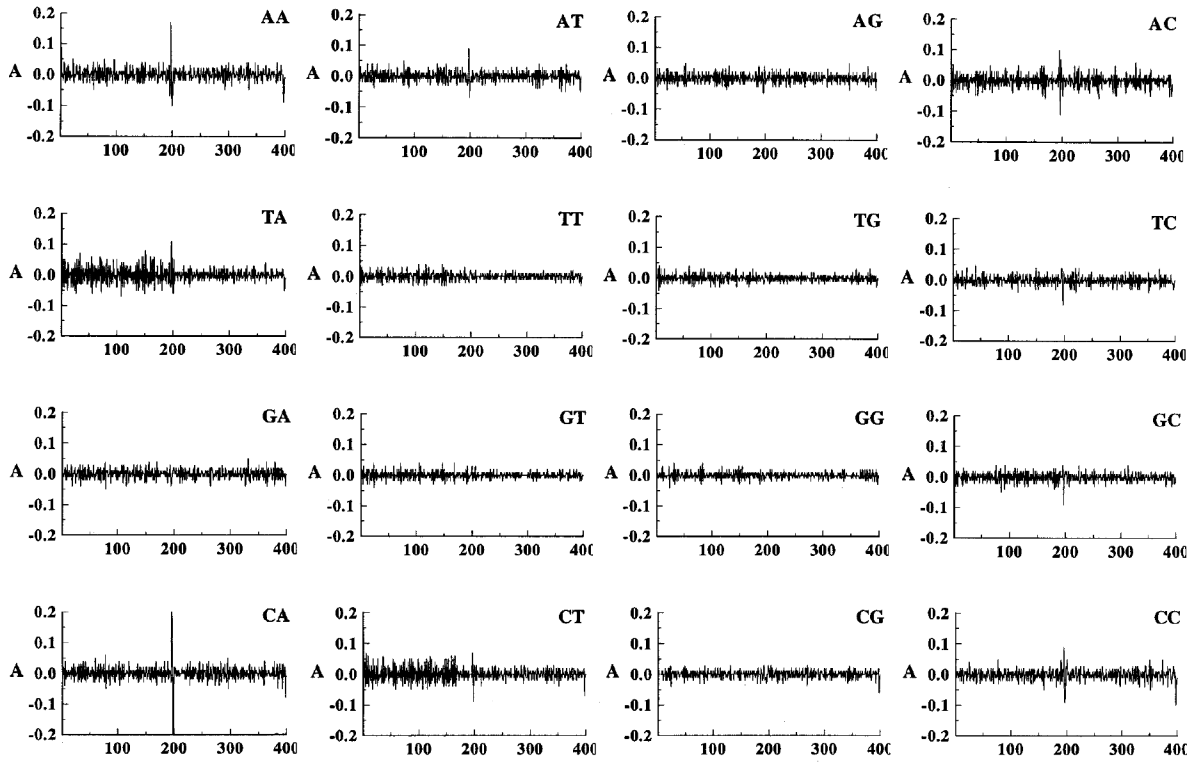

Figure 12. Plot of NDF gradient as a function of nucleotide position from 200 nucleotides prior to TAG up to 200 nucleotides post-TAG for the genome of *S. cerevisiae*.

presumably because the composition of the coding regions has to conform to constraints imposed by the amino acid composition of genes as well as the DNA structure, whereas the constraints of the non-coding (i.e. pre-ATG) regions are likely to be purely structural. In accordance with this, we observed that the standard deviations in the pre-ATG regions were rather low (~ 0.01–0.02).

A list of the minimum and maximum ratios in variances is shown in Table 1. As is obvious from the values, ratios that are very different from 1 indicate that the variances change most markedly at the boundary. Consistent with our previously described results, this happens for one of AC, TA or CT in all cases. This clearly shows that the appearance of the gradient signal is also reflected in the statistical properties of the density function. For three cases, the change in variance is most pronounced for CG. Although this is not reflected in the gradient signal for any of the genomes, it could be a complex outcome of the CG contents of the non-coding and coding regions that are known to be very different. Further, our observation that the variation is not seen for CG alone suggests that the signal is unlikely to be a simple function of base composition.

The absence of a simple correlation between the dinucleotide counts (or frequencies) and the distribution function described here suggests that the gradient is a characteristic of the interface and cannot be deduced from the overall composition of the sequence alone. The structure of DNA is determined at the primary level by the dinucleotide[14]. The change in the distribution of some dinucleotides at the boundary discussed in this report, is likely to be due to a change in DNA structure that is recognized by the read-out apparatus. It is difficult to envision a one-step mechanism that effects this; indeed, such a mechanism would probably be too fragile. A more probable scenario is through subtle changes of the DNA structure, through stacking interactions. The nature of the 3′ nucleotide is known to determine the energy of stacking and consequently the DNA stability locally[15,16]. Further, the region that is proximal to the beginning of a gene can be taken to possess a lower stability, as it would lead to facile opening of the two strands. Consistent with this, we observe that the most stable stack 5′GC3′ does not show a signal, in spite of the fact that GC-richness is one of the features that distinguishes coding and non-coding regions[5]. The 5′TA3′ stack on the contrary, is the least stable one and in our analysis is found to exhibit a maxima at the coding and non-coding interfaces in all the genomes.

Macroscopic curvature in DNA has been shown to be of immense biological significance[17]. Indeed, regions upstream of genes are also known to be intrinsically curved[18,19]. The marked change in the NDF at the interface may also be a consequence of increased DNA flexibility in the regions before the start codon. A finer analysis of the pre-ATG region, through the refinement of the nucleotide distribution function is likely to provide important clues to the location of promoters and other regulatory elements.

The universality of the signals across genomes, hints at certain basic structural characteristics of genes. The property we have investigated expresses an overall (in a statistical sense) property of a genome. In an individual sequence there are signals of the gradient (of comparable magnitudes) at a number of points, including the ATG position. However, except at the ATG positions, the signals are not preserved. Thus a statistical averaging over a number of sequences reduces the signals over the non-ATG sites to negligible values. Therefore, the gradient measure, in its present form, cannot be claimed to be a gene-finding algorithm.

1. Bibb, M. J., Findlay, P. R. and Johnson, M. W., *Gene*, 1984, **30**, 157–166.
2. Zhang, C. T. and Zhan, Y., *J. Theor. Biol.*, 1994, **167**, 161–166.
3. Bucher, P., Fickett, J. W. and Hatzigeorgiou, A., *Comput. Appl. Biosci.*, 1996, **12**, 361–362.
4. Almirantis, Y., *J. Theor. Biol.*, 1999, **196**, 297–308.
5. Fickett, J. W. and Tung, C. S., *Nucleic Acids Res.*, 1992, **20**, 6441–6450.
6. Delcher, A. L., Harmon, D., Kasif, S., White, O. and Salzberg, S. L., *Nucleic Acids Res.*, 1999, **127**, 4636–4641.
7. Blattner, F. R. *et al.*, *Science,* 1997, **277**, 1453–1474.
8. Bult, C. J. *et al.*, *Science,* 1996, **273**, 1058–1073.
9. Smith, D. R. *et al.*, *J. Bacteriol.*, 1997, **179**, 7135–7155.
10. Goffeau, A. *et al.*, *Science,* 1996, **274**, 563–567.
11. Fraser, C. M. *et al.*, *Nature*, 1997, **390**, 580–586.
12. Cole, S. T. *et al.*, *Nature*, 1998, **393**, 537–544.
13. Nussinov, R., Sarai, A., Smythers, G. W. and Jernigan, R. L., *Biochim. Biophys. Acta*, 1989, **1008**, 329–338.
14. Calladine, C. R. and Drew, H. R., *J. Mol. Biol.*, 1984, **178**, 773–782.
15. Friedman, R. A. and Honig, B., *Biophys. J.*, 1995, **69**, 1528–1535.
16. Gupta, G. and Sasisekharan, V., *Nucleic Acids Res.*, 1978, **5**, 1639–1653.
17. Harrington, R. E., *Mol. Microbiol.*, 1992, **6**, 2549–2555.
18. Nickerson, C. A. and Achberger, E. C., *J. Bacteriol.*, 1995, **177**, 5756–5761.
19. Schatz, T. and Langowski, J., *J. Biomol. Struct. Dyn.*, 1997, **15**, 265–275.

ACKNOWLEDGEMENTS. We thank Prof. S. K. Brahmachari, Functional Genomics Unit, Centre for Biochemical Technology, New Delhi and Prof. N.V. Joshi, Centre for Ecological Sciences, Indian Institute of Science, Bangalore for critical comments and helpful discussions on the work presented here.

Received 4 April 2001; revised accepted 28 August 2001