STATISTICAL CHARACTERIZATION OF CONSERVED NON-CODING ELEMENTS IN VERTEBRATES

te Boekhorst R.1, Walter K.2, Elgar G.3, Gilks W.R.2, Abnizova I.2
1 University of Hertfordshire, College Lane, Hatfield, UK; 2 MRC-BSU, Robinson Way, Cambridge, UK; 3 Queen’s Mary College, London, UK

* Corresponding author: e-mail: r.teboekhorst@herts.ac.uk

Key words: CNE-elements, non-coding DNA, regulatory DNA

SUMMARY

Motivation: Recently, a set of highly Conserved Non-coding Elements (CNE’s) was derived from a Fugu-human genome comparison. We characterise some statistical features common to these elements in order to facilitate their identification in silico.

Results: We found a pronounced pattern around the borders of CNEs: GC-rich flanking regions of low entropy compared to AT-rich, high entropy sequences within the borders. We also identified the most abundant significant motifs inside and adjacent to the borders of CNE’s. At the borders, motifs are significantly clustered which points to their possible role as binding sites.

INTRODUCTION

Only around 1.2 % of human DNA is known to be coding for proteins. Our knowledge of the role and location of other elements is limited and new types of sequences of unknown function are still discovered. Recently, several sets of highly conserved non-coding sequences have been identified in vertebrate genomes (Woolfe et al., 2004; Bofelli et al., 2005; Dermitzakis et al., 2005). A combination of comparative genomic studies and laboratory experiments has shown that these conserved non-coding elements (CNEs), most of which are more conserved than protein-coding exons, may be regulatory elements (Moses et al., 2005; Xie et al., 2005).

Conserved regulatory regions have been the objects for motif discovery by phylogenetic foot-printing algorithms (Blanchette, Nompa, 2002). However, most efforts been related to the promoter motifs (FitzGerald et al., 2004) and although CNEs appear to have striking “signatures” (Walter et al., 2005), little motif discovery has been done for CNEs.

Here, we focus on the motif identification and statistical characterization of the CNEs collected by Woolfe et al. (2004 ). Based on a MEGABLAST comparison between human and pufferfish (Fugu ribripes) genomes, they identified about 1400 highly conserved non-coding sequences. Most of these sequences are located in and around developmental regulation genes and when some of them were tested in the laboratory, they appeared to drive tissue-specific gene expression in early development (Woolfe et al., 2004). These facts encouraged us to consider CNEs as putative regulatory regions, namely enhancers, and to check whether they could be characterised by some of the basic statistical properties of regulatory regions such as the abundance (Papatsenko et al., 2002) and the typical spatial distribution (FitzGerald et al., 2004) of binding motifs.
MATERIALS AND METHODS

The set of CNEs identified by Woolfe et al. (2004) contains 1373 elements, vary in size from 53 bp to 740 bp (mean length 200 bp) and a level of conservation is from 68 % to 98 % identity. We use this data set to build up a CNE lexicon, and check for presence of statistical properties typical of cis-regulatory regions. To characterise CNE borders, we generated two positive data sets of 5′ and 3′ CNE flanking regions of 50 bp each (upCNE and CNEdown) from the 1231 CNEs which are longer than 100 bp.

**Likelihood of motifs.** It is known that certain sequences that operate as “binding motifs” are surprisingly abundant within regulatory regions. Given the DNA composition of a region, the globally most abundant motifs may be defined as those that are most likely to occur. In the work reported here, we determined the motifs in the CNE-flanking alignments with the highest likelihood. We will show that these motifs are more abundant than expected due to the composition of the upCNE and downCNE regions. To do so, we generated a large number (10000) of “surrogate” alignments with the same position-dependent composition as the CNE alignment under consideration by randomly shuffling the original sequence 10000 times. Next, Z-scores of words from the original lexicon were calculated as standardized deviations from the mean frequency (of the same words) of the randomised alignments. Words from the original CNE alignment are defined as significant if their Z-score exceeds 2 standard deviations.

**Spatial Distribution (clustering) of Words.** Some words, not necessarily the most frequent in the CNE alignments, could be functionally important as binding sites and could therefore be clustered around CNE borders. We use the local frequency of words (i.e. within columns) in the alignment to determine their degree of clustering. To assess the statistical significance of word clustering, the clustering coefficient, CC, is defined for each word $x_i$ in each start position $j$ and sequence (we omit the indexing of sequence for simplicity) as $CC_j(x_i) = \frac{N^j(x_i) - \bar{N}(x_i)}{\sigma(x_i)}$, where $N^j(x_i)$ is the occurrence of word $x_i$ starting in position $j$, $\bar{N}(x_i)$ is the mean frequency (i.e. of $N^j(x_i)$) over all positions $j$ in the alignment, and $\sigma(x_i)$ is the standard deviation of $N^j(x_i)$. A word $x_i$ is significantly locally clustered (or anti-clustered) in position $j$, if $|CC_j(x_i)| > 2$.

**Compositional homogeneity.** While aligning CNEs, we had the impression that the flanking regions vary stronger in composition than the CNEs themselves. We estimated the di-nucleotide entropy separately in each 50 bp flanking region and each 50 bp CNE sequence of the alignments as a quantitative measure for their compositional diversity. In previous work, we have shown that the entropy of regulatory regions is intermediate between that of coding (highest entropy) and non-coding, non-regulatory regions (lowest entropy) (Orlov et al., 2006).

**Sequential persistence.** We used the Hurst exponent as a measure of the stationarity of the DNA sequence around the CNEs border. The Hurst coefficient was calculated by Rescaled Range Analysis. We applied this method by transforming a DNA sequence into a binary code of $x_k = +1$ for $k = G, C$ and $x_k = -1$ for $k = A, T$ (Orlov et al., 2006). In case of random, identical and independent occurrences of nucleotides in DNA, H equals 0.5. A high Hurst exponent (> 0.5) points to extensive autocorrelations (i.e. non-stationarity). A series that contains a significant change in composition is therefore expected to be characterized by H > 0.5. In previous work, we have shown that the Hurst exponent of regulatory regions ($H \sim 0.62$) is intermediate between that of coding ($H < 0.5$, indicating anti-persistence) and non-coding, non-regulatory regions ($H \sim 0.67$) (Orlov et al., 2006).

To statistically characterise the CNEs and their borders, we calculated Entropy and Hurst exponent between the following regions. For entropy: 50bp upstream flanking...
regions, the first- and last 50 base pairs of the CNE itself and 50 bp downstream flanking regions. The entropy values of these stretches were compared to sequences (50 bp long) of randomly picked non-coding, non-regulatory DNA in *Fugu*. For the Hurst exponent: an upper CNE bordering region containing the first 50 base pairs before and after the CNE start position and a lower CNE bordering region consisting of the last 50 base pairs of a CNE and the first 50 base pairs after the stop-position of that CNE. We compared Hurst exponent values of the two bordering regions with those of randomly selected non-coding, non-regulatory DNA within the same window length as the aligned CNEs.

**RESULTS**

We aligned the highest scoring significant words (of length 12 here) with respect to their start positions in columns, and put them into to WebLogo format (Crooks *et al.*, 2004). The “logos” of these over-represented words are shown in Fig. 1. The lowest scoring words are visualized in Fig. 2. As one can see from Fig. 1–2, over- and under-represented words are AT rich and CG poor respectively and their start positions appear to cluster close to the CNE borders.

![Figure 1.](image1.png)

*Figure 1.* Left: highest scoring 12-mers in CNE<sub>down</sub>, the CNE border is at position 50. Right: high scoring 12-mers in upCNE, the CNE border is at position 51.

![Figure 2.](image2.png)

*Figure 2.* Left: lowest scoring 12-mers in CNE<sub>down</sub>, the CNE border is at position 50. Right: lowest scored 6- and 12-mers in upCNE, the CNE border is at position 51.

We found that there this clustering (CC > 2) of certain short patterns around the CNE borders is significant. This is shown in the plot of start-positions of significantly large clustering coefficients (> 2) along the alignment (Fig. 3).

The average entropy of upstream and downstream flanking regions (both E = 2.43) is significantly lower than that of regions within the CNE border (both E = 2.52) (Newmans-Keuls ad hoc comparisons after a 1 way repeated measurement ANOVA within sequences. A Friedman test – as a non-parametric alternative to the repeated measurement ANOVA – backed up the results). The entropy of non-coding, non-regulatory regions (E = 2.49) is significantly higher than that averaged over the two flanking regions (t-Test: t = 8.52, df = 2460, p < 0.0001) and significantly lower than that averaged over the two within CNE regions (t-Test: t = 5.82, df = 2460, p < 0.0001). Mann-Whitney U tests, as non-parametric alternatives to the parametrical t-Tests demonstrated a significant difference between the non-coding, non-regulatory DNA on the one hand and the combined (= averaged) data of the two flanking regions on the other hand, but not with the combined (averaged) within CNE sequences.
Figure 3. The clustering of four mers around the upCNE border, at position 50 bp. The vertical axis shows modified CC (= sign(CC)*CC2) for visualization purposes. The horizontal axis shows position in the upCNE alignment.

The Hurst exponents between upper and lower bordering regions (respectively $H = 0.65$ and $0.66$) do not differ significantly, but do so between the values of each of the two bordering regions and those of non-coding, non-regulatory regions ($H = 0.59$) (Newmans-Keuls ad hoc comparisons after a 1 way ANOVA on log-transformed data). Together, these results point to a change in composition at the borders of CNEs.

DISCUSSION AND CONCLUSION

We have showed that the motifs around CNE borders are not just the consequence of compositional bias. In addition, we identified the following statistical “signatures” of CNEs: (i) the sequences around CNE borders are surprising rich in globally and locally over-represented motifs; (ii) CNE borders appear to correspond to a pronounced change-point in composition; (iii) flanking CNE sequences have low entropy and are CG rich whereas the CNE themselves are AT-rich, and have a higher entropy compared to the flanking regions. Although it has been put forward that some CNEs might be matrix attachment regions (Glazko et al., 2003) or participate in inter-chromosomal interactions (Muller, Schaffner, 1990), due to their statistical properties CNEs might indeed function as regulatory regions. They contain more statistically significant abundant words than expected by chance, many words are clustered close to their start positions and their entropy is on average 2.52, i.e. in between that of typical coding regions (~2.68, refs) and non-coding, non-regulatory regions (2.48). These findings corroborate evidence in the literature (Hardison, 2000; Nobrega et al., 2003). The most significantly clustered motifs could be candidates for TFBS cores. Note that some words are rarer than expected (anti-clustered) near the border (Fig. 3). They could be candidates for under-represented TFBS binding sites. Comparison with randomly picked non-coding non-regulatory (NCNR) DNA revealed that in the latter locally
highly clustered motifs are fairly uniformly spread over alignments, in contrast to the clustering around the borders of CNEs, which is typical for regulatory regions (FitzGerald et al., 2004).

REFERENCES


