Widespread DNA Structural Constraint in the Human Genome

Stephen C.J. Parker[1], Loren Hansen[1,2], Eric Bishop[1], David Landsman[2], NISC Comparative Sequencing Program[3], the ENCODE[4] Multi-species Sequence Analysis (MSA) Group, Elliott H. Margulies[3], and Thomas D. Tullius[1,5]

[1]Bioinformatics and Systems Biology Program, Boston University, Boston, Massachusetts
[2]National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland, USA
[3]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA
[4]ENCyclopedia Of DNA Elements Project Consortium (www.genome.gov/encode)
[5]Department of Chemistry, Boston University, Boston, Massachusetts

INTRODUCTION

Much of the success in deciphering functional signals in the vast non-coding landscape of the human genome has come from advances in detecting sequences that are conserved between multiple species. However, fundamental questions still exist that are recalcitrant to traditional genomic analysis. For example, results from the ENCODE Pilot Project show that many functional elements are not evolutionarily constrained at the primary sequence level. We have developed a novel algorithm to assess evolutionary constraint based on DNA structure (as measured by the hydroxyl radical cleavage pattern) rather than primary sequence.

After correcting for false discovery rates, we find that about twice as much territory in the ENCODE regions of the human genome is covered when the set of structurally-constrained regions are merged with the set of sequence-constrained regions. Furthermore, regions that are identified as constrained based on structure and not primary sequence are not distributed at random. Instead, these regions overlap functional elements (e.g., DNase I hypersensitive sites, promoters, enhancers, histone modification sites). That is, some functional elements that are not constrained at the level of primary sequence are constrained based on DNA structure. Surprisingly, the majority of functional annotations conform to this paradigm: a larger proportion of elements are constrained based on structure compared to sequence alone. This increased enrichment is significantly more than what is expected at random.

HYDROXYL RADICAL CLEAVAGE PATTERNS OF DNA

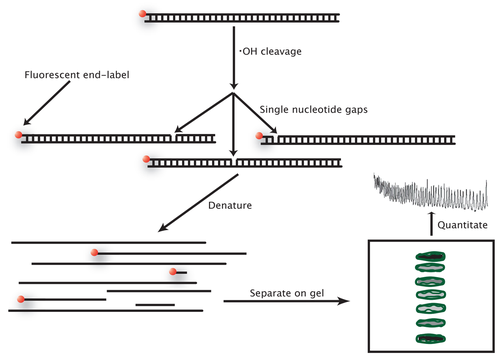


Figure 1. ***Cleavage schematic*** *- First, labeled DNA is subjected to hydroxyl radical cleavage, which should produce single nucleoside gaps at every position in the strand. Next, the DNA fragments are run out on a denaturing gel. Each successive band in the gel represents a DNA fragment that differs in length by one nucleotide. The intensity of each band represents the amount of cleavage at that particular nucleotide. Cleavage patterns are a representation of the local structural profile of a DNA molecule.*

DIFFERENT DNA SEQUENCES WITH SIMILAR CLEAVAGE PATTERNS

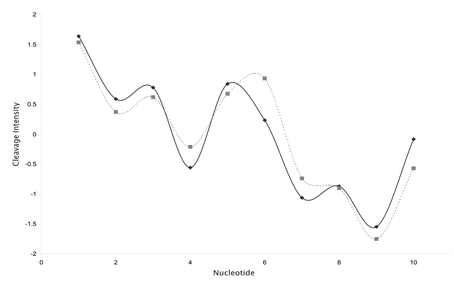


Figure 2. ***Low sequence identity 10mers with high cleavage similarity*** *- The hydroxyl radical cleavage patterns of two 10mer DNA sequences with 0% identity are plotted. A Pearson correlation of 0.94 indicates a highly similar pattern.*

PREDICTING CLEAVAGE PATTERNS

Several algorithms were tested for their ability to accurately predict the amount of hydroxyl radical cleavage at each nucleotide. The algorithm that has produced the most accurate results is a sliding tetramer window (STW) algorithm. It functions by dividing the sequence into overlapping tetramers and retrieving the cleavage intensity information for each tetramer from the database. It then averages across the length of the sequence to yield a prediction. Figure 3 shows a comparison of the predicted and experimentally-determined patterns. The correlation between predicted and experimentally determined cleavage patterns is strong, with an average Pearson correlation coefficient of ~0.85.

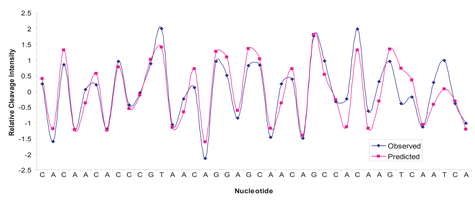
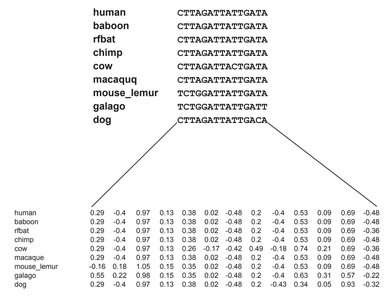


Figure 3. ***Comparison of predicted and experimentally determined cleavage pattern*** *- Using our Sliding Tetramer Window (STW) algorithm, the cleavage intensity of a particular 40mer sequence was predicted. The results are plotted here, along with the experimentally determined cleavage intensity of the same DNA fragment.*

STRUCTURE CONSERVATION ALGORITHM

Because different DNA sequences can have similar structures (Figure 2) and local DNA structural motifs are known to play a role in some biological functions, there maybe be segments of a genome that are conserved based on structure and not sequence. Accordingly, we sought to develop an algorithm to measure the conservation of DNA structure. Starting with an existing multiple alignment, we then predict the hydroxyl radical cleavage pattern for each sequence (Figure 4). We next implement a modified version of the binCons algorithm whereby Euclidean distance is used to measure the similarity between the predicted cleavage pattern vectors across different species (see equation below).



human	0.29	-0.4	0.97	0.13	0.38	0.02	-0.48	0.2	-0.4	0.53	0.09	0.69	-0.48
baboon	0.29	-0.4	0.97	0.13	0.38	0.02	-0.48	0.2	-0.4	0.53	0.09	0.69	-0.48
rfbat	0.29	-0.4	0.97	0.13	0.38	0.02	-0.48	0.2	-0.4	0.53	0.09	0.69	-0.36
chimp	0.29	-0.4	0.97	0.13	0.38	0.02	-0.48	0.2	-0.4	0.53	0.09	0.69	-0.48
cow	0.29	-0.4	0.97	0.13	0.26	-0.17	-0.42	0.49	-0.18	0.74	0.21	0.69	-0.36
macaque	0.29	-0.4	0.97	0.13	0.38	0.02	-0.48	0.2	-0.4	0.53	0.09	0.69	-0.48
mouse_lemur	-0.16	0.18	1.05	0.15	0.35	0.02	-0.48	0.2	-0.4	0.53	0.09	0.69	-0.48
galago	0.55	0.22	0.98	0.15	0.35	0.02	-0.48	0.2	-0.4	0.63	0.31	0.57	-0.22
dog	0.29	-0.4	0.97	0.13	0.38	0.02	-0.48	0.2	-0.43	0.34	0.05	0.93	-0.32

Figure 4. ***Initial step of structure conservation algorithm*** *- Using our STW algorithm, the hydroxyl radical cleavage patterns for all sequences in an existing multiple alignment are predicted. This process was done for 1% of the human genome (the ENCODE regions).*

$$S(k) = \frac{\sum_{y=1}^{N-1}\left(-\log\sqrt{\sum_{j=1}^{n}(h_j - y_j)^2}\right) \times w_y}{N-1}$$

Where:
S = Structure score for window k
N = number of species in alignment
n = window size
j = position in window
h_j = structure value of human sequence at position j
y_j = structure value of species y sequence at position j
w_y = weight based on phylogenetic distance of species y to human

QUANTIFYING THE AMOUNT OF STRUCTURAL CONSTRAINT

We generated null alignments to estimate the amount of structural constraint in real alignments. All results are reported at a threshold that represents a 5% False Discovery Rate.

	W=7	W=9	W=11	W=13	W=15	W=17	W=19
K=2	5.8(10.4)	9.6(12.9)	13.8(15.5)	19.9(16.8)	21.6(17.6)	23.2(18.4)	29.3(18.2)
K=3	2.6(5.4)	4.1(6.8)	6.3(8.2)	9.2(9.3)	11.6(11.3)	14.1(12.2)	17.6(11.8)
K=4	NA	2.1(4.3)	3.8(5.4)	5.2(6.1)	6.5(7.1)	7.7(8.6)	9.7(8.9)
K=5	NA	0.8(2.9)	2.2(3.8)	3.3(4.2)	4.5(5.1)	5.9(6.1)	5.7(6.4)
K=6	NA	0.5(2.3)	1.0(2.8)	2.2(3.5)	3.2(4.0)	3.9(4.5)	4.7(5.1)
K=7	NA	0.3(1.5)	0.6(2.0)	0.9(2.6)	2.0(3.1)	3.2(3.4)	3.7(4.2)

Table 1. ***Comparison of the percent of ENCODE covered by sequence (blue) and structure (red) constrained regions*** *- The binCons algorithm is compared directly to our structure conservation algorithm by using different window (W) sizes. Neutral alignments are generated by bootstrapping multiple columns (K) from an existing alignment. All values represent coverage at a 5% False Discovery Rate (FDR). Parameter combinations that yield unreasonable sequence coverage compared to prior studies are grayed out. Underlined values represent the mean from ten trials.*

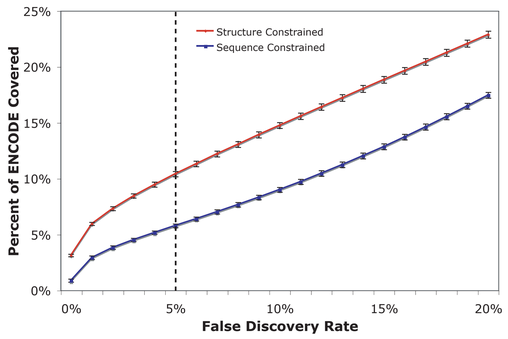


Figure 5. ***Percent of ENCODE covered at different False Discovery Rates*** *- The amount of constrained territory is higher for structure compared to sequence at a given FDR. At a 5% FDR 10.4% of ENCODE is constrained based on structure while 5.8% is constrained based on sequence. The above data reflects the binCons and structure conservation algorithm run with a window (W) size of 7 and with null alignments generated by bootstrapping 2 (K) columns at a time from an existing alignment. Error bars represent one standard deviation from the mean of ten trials.*

DISTRIBUTION AND LOCATION OF STRUCTURAL CONSTRAINT

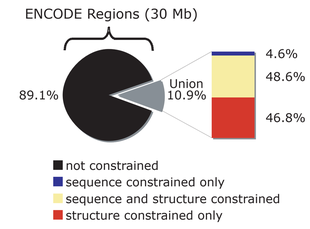


Figure 6. ***Distribution of sequence and structure constraint in ENCODE*** *- When structure constrained regions are merged with sequence constrained regions about twice as much territory is covered within the ENCODE regions.*

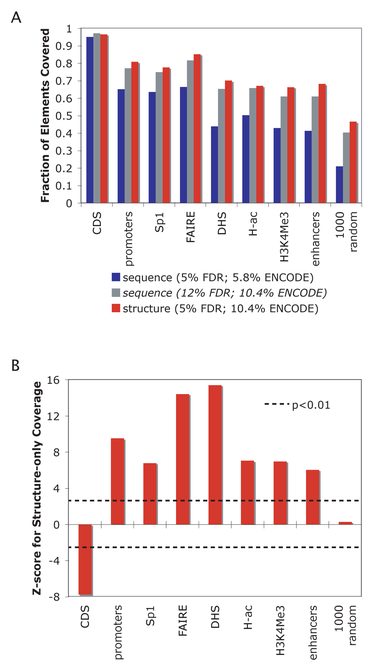


Figure 7. ***Structure constrained territory preferentially occurs in functional elements*** *- More functional elements are covered by structure constrained regions compared to sequence constrained regions (A). Structure constrained elements overlap with functional elements significantly more than random expectation with the exception of CDSs where they overlap significantly less (B). Z-scores are based on 1,000 randomizations.*

CONCLUSION

Our new high-resolution DNA structure conservation method reveals that structural constraint is widespread throughout the human genome, and that these regions are informative of known functional sites. That natural selection operates to preserve not only information encoded in the sequence of DNA, but also in its local structure, may be of critical importance to understanding how the human genome functions, and to refining what is meant by 'constrained sequence.'

SUMMARY

- We have organized a database of hydroxyl radical cleavage patterns of DNA and are actively collecting data.
- We are able to predict the cleavage pattern of any DNA sequence with high accuracy.
- The distribution of evolutionary constrained cleavage patterns is much larger compared to primary DNA sequence constraint.
- The additional territory covered by structurally constrained regions occurs specifically in experimentally annotated functional elements.