Evolutionary Constraint of DNA Structure in Non-Coding Functional Elements

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

[1]Genome Technology Branch and [2]National Center for Biotechnology Information, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA.
[3]Program in Bioinformatics and [4]Department of Chemistry, Boston University, Boston, Massachusetts, USA
[5]ENCyclopedia Of DNA Elements Project Consortium

ABSTRACT

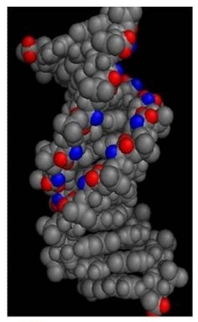
Recent analyses have revealed that large portions of many classes of non-coding experimental annotations identified by the ENCODE consortium show no evidence of primary sequence constraint during mammalian evolution. Since the current set of evolutionarily constrained regions only represents sequence whose primary order of nucleotide bases have remained highly similar throughout evolution, we sought to explore other properties of DNA that might be evolutionarily constrained and thus important for function; one such property is DNA structure. Importantly, it has been shown that different primary sequences exhibit very similar 3-dimensional structure as assessed by hydroxyl-radical cleavage patterns.

We used this information to develop a "structure" conservation score that assesses the similarity of structure from multi-sequence alignments. Using very stringent false discovery rates, we can identify at least twice as much sequence being constrained in structure, compared to primary sequence constraint methods. Half of these new structurally-constrained regions are not evolutionarily constrained at the primary sequence level. Interestingly, we observed a better correlation between structural constraint and many non-coding functional annotations, than with primary sequence constraint. Indeed, some functional annotations have up to twice as much structural constraint compared to primary sequence constraint.

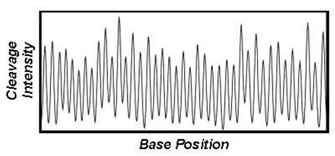
Our results support the hypotheses that 1) other more complex functional "grammars" likely exist in the genome, 2) they are important for genome function, and 3) such structural grammars appear to be evolutionarily constrained and can be detected by comparative sequence analysis.

Is DNA Structure Evolutionarily Constrained?

Hydroxyl Radical Cleavage Intensity Correlates with Solvent Accessibility



Blue: 4' hydrogens
Red: 5' hydrogens



Greenbaum, Pang and Tullius, Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Research*, (2007) In Press

Structure = Solvent Accessibility

- Different primary sequences can have similar structures
- Structure can be computationally predicted from primary sequence

Generating a Structure Conservation Score

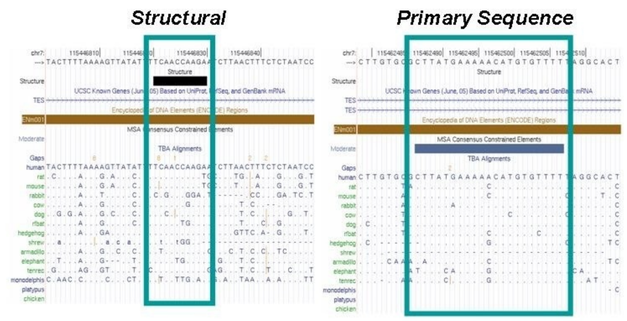
```
ACTTACTTATCTAATGAAAAGTGCCCAGCATAAAAATGCAGG human
--TTACTTATCTAATGAAAAGTGCCCAGCATAAAAATGCAGG chimp
ACTTACTTATCTAATGAAAGGTGCCCAGCATAAAAATGCAGG baboon
--TTACTTATCTAATGAAAGGTGCCCAGCATAAAAATGCAGG macaque
ATTTACTTATCTAATCAAAGGTGTCCAGTATAAAAATGCAGG marmoset
-CTGAATTATCTAATCAAATATGCCCTGCACAAAAATTCAGG galago
---TAGTTCTTTAATCACAGGTGCTAAGCGTAAAACTTCACG rat
--TTATTTCTTTAATCACAGGTGCTAAGCATAAAACTTCAAG mouse
ACTTACTTCTCTAATCAGAGGTGCTCAGTATATGAATTCAGG cow
ACTTACTTCTCTAATCAAAGGTACTCAGCATAAAAATTCAGG dog
ACTTTCTTTTCTAATCAAAGGTACACAGTACAAGAATCCAGA shrew
A-T-ACTTCTCTATTCCAAGGTGCCTCGCATGAAAATTCAGG armadillo
```

```
-0.02  0.29 -0.27  0.62  0.02 -0.09 -0.13 -0.08  0.21  0.43 human
-0.02  0.29 -0.27  0.62  0.02 -0.09 -0.13 -0.08  0.21  0.43 chimp
-0.02  0.29 -0.27  0.62  0.02 -0.09 -0.13 -0.08  0.21  0.43 baboon
-0.02  0.29 -0.27  0.62  0.02 -0.09 -0.13 -0.08  0.21  0.43 macaque
-0.02  0.29 -0.27  0.62  0.02 -0.09 -0.14 -0.01  0.35  0.06 marmoset
 0.14 -0.13 -0.40  0.62  0.02 -0.09 -0.14 -0.01  0.35  0.06 galago
 0.75 -0.05 -0.16 -0.06  0.21  0.00 -0.43 -0.08  0.06 -0.15 rat
-0.19 -0.37 -0.32 -0.06  0.21  0.00 -0.43 -0.08  0.06 -0.15 mouse
 0.09  0.18  0.03  0.01  0.28  0.10 -0.14 -0.01  0.32  0.13 cow
 0.09  0.18  0.03  0.01  0.28  0.10 -0.14 -0.01  0.35  0.06 dog
-0.05  0.32 -0.14 -0.20 -0.23 -0.08 -0.14 -0.01  0.35  0.06 shrew
-0.01  0.18  0.03  0.01  0.37  0.24 -0.07  0.93  0.09 -0.14 armadillo
```

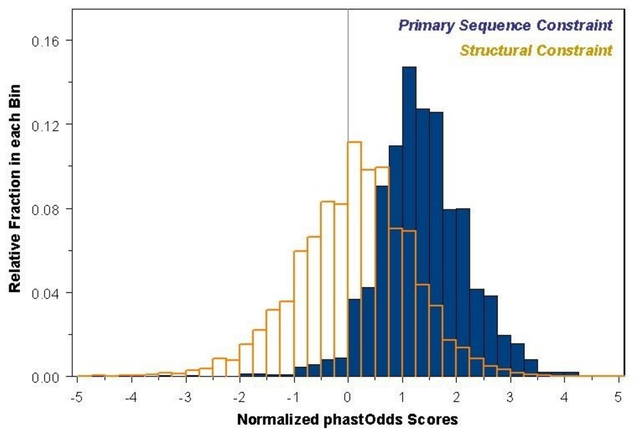
Similarity Score Based on Euclidean Distance

Properties of Structurally Constrained Regions

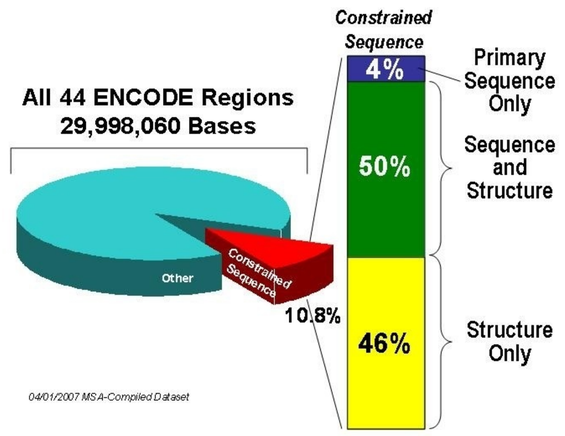
Example of Structural Constraint *vs.* Primary Sequence Constraint



Primary Sequence Conservation of Structurally Constrained Regions

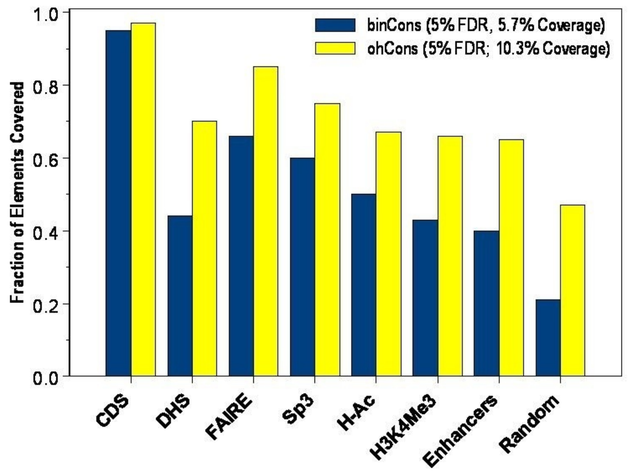


Primary and Structural Sequence Constraint

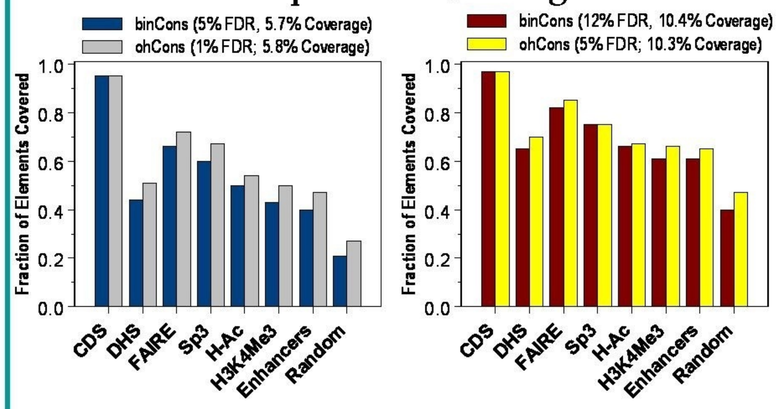


Correlation with Functional Annotations

Equivalent False Discovery Rate



Equivalent Coverage



Significance of Structure-Only Overlap