“Scaling of Sequence: Impact of Interactome?”

Jonathan Miller[1]

“I have called this principle, by which each slight variation, if useful, is preserved, by the term Natural Selection.” **- Charles Darwin, Origin of Species. Evidence has recently been presented for strong spatial correlations of genome sequence conservation. These correlations scale, suggesting a framework for incorporating them into estimates of the significance of genomic sequence conservation, where in the past only phylogeny and site-specific substitution rates have been accounted for. Here, new examples of these phenomena, including “ultraconservation” of amino-acid sequences, are presented.**

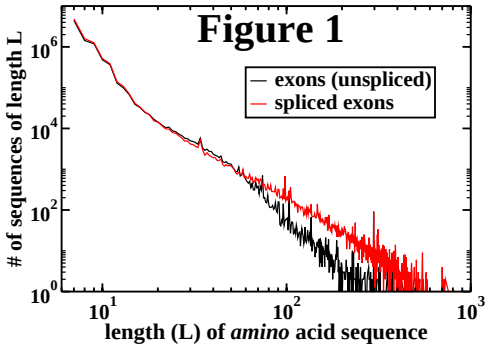
Keywords **— comparative genomics, evolution, ultra-conservation, sequence correlations, scaling.**

I. Purpose

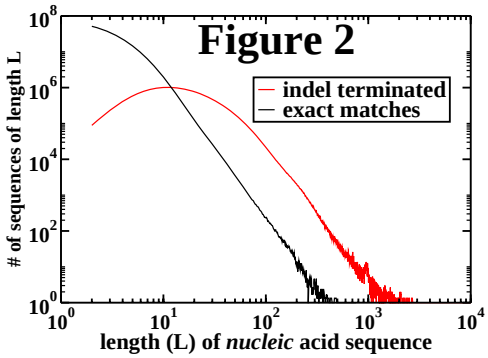
THE catalog of biological signal transducers and genetic regulatory elements has expanded rapidly in recent years, and now includes DNA motifs, proteins, and a rich parallel universe of structural, enzymatic, messenger, and regulatory RNAs. Although acting by distinct mechanisms, all of these elements are created equal: they are digitally encoded within the genome. Elements can’t interact unless co-expressed, and the spatial organization of the genome is a determinant of intensity and timing of expression. In turn, genomic sequence reflects the primary, secondary, tertiary and quaternary features of these elements and their interactions.

Recently, hints have emerged of how some of these interactions are built into a genome. Perfect conservation of long runs of genomic sequence among diverse species is proposed to arise from multiple, non-local and overlapping interactions of gene products, so that to retain function, any mutation in such an element must be complemented by multiple further mutations in sequences throughout the genome. It was subsequently observed that perfectly-conserved genomic sequence has an algebraic length distribution, with spatial locations along the genome strongly correlated [1].

Here I report two novel and related observations in human and mouse: (i) Perfectly-conserved protein (amino-acid) sequences - a less stringent constraint on genomic sequence than base identity - also exhibit an algebraic length distribution (figure 1). Some of these sequences span coding exons, suggesting that their conservation reflects selection on protein function.



(ii) Imperfect - but indel-free - genomic sequence conservation exhibits much the same exponent as perfectly-conserved sequence, but at greater scales (figure 2).



Both of these phenomena are generic to metazoans. It is proposed that they arise from (i) a combination of neutral *duplication* and mutation that constantly produces new candidate sequences; followed by (ii) fixation or loss of the candidates. Thus, a new genomic element can arise without disrupting the multiple, non-local and overlapping network of interactions among gene products that enforce perfect conservation. Nevertheless, a novel altered network is created - coexisting in the short term with the existing network - whose impact on the fitness of the organism will determine ultimate fixation or loss of the novel sequence element.

II. Conclusion

It is hypothesized that scaling of conserved sequence reflects a modular, hierarchical structure of genomes and their encoded interactomes.

References

[1] Salerno W, Havlak P, Miller J (2006). Scale-invariant structure of whole-genome intersections and alignments. *Proc Natl Acad Sci* USA, 103(35): p. 13121-5.

Acknowledgements: The author wishes to thank Richard A. Gibbs for advice and support.

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston TX USA. E-mail: jnthnmllr@gmail.com