**Non random Mutation:**

**Nature 4, 2008. Single-nucleotide mutation rate increases close to insertion/deletion in eukaryotes. 455:pg 105-108**

**The Regional Difference hypothosis holds that variations in mutation rates are due to structural diversity and selection pressure. But the authors of this paper find that single base substitution rate increases in the 200 bp region surrounding instertion/deletion points, suggesting a casual relationship between one form of mutation and another.**

**mutational heteroginisity is known to occur within and between species and at multiple physical scale.**

**Silva, Kondrashov, Trends Genet 18, 544-547; Ellegren, Webster. Mutation rate variation in the mamallian genome. Curr Opin Genet Dev; Denver and Thomas. High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. Nature; Longman-Jacobsen, Gaudieri. Gene 2003.**

[**Properties of GA-Sequences**](http://www.plosone.org/article/info:doi/10.1371/journal.pone.0003818)

Albrecht-Buehler G PLoS ONE 2008

Summary:

Guenter looks at the pure GA sequences larger than 50 basepairs which are stastically improbable, but present in a Pareto distribution (non random) by the tens of thousands in humans and chimpanzees and at 4-5 fold higher densities in rodents. Guenter argues that diffusion is too slow to guide transcript factors to their targets (think heat shock) and that navigation markers must be present genome. He suggests that pure GA sequences may be these navigation markers.

Glossary:

Pareto-distributions wealth, sand mass, and city population size

Pareto-equation f(x, k, x.min) = ( k \* x.min^k ) / x^(k + 1)

Pareto v Poisson higher large value frequencies (pareto)

Poisson equation    f(k, o) = ( o^k \* e^-o ) / ( k! )

Poisson distribution explains the distribution of radioactively induced mutations

hexanucleotide spectrum    Distribution of 6bp amino acids

Inspirations:

Pareto distributions describe things that interact to combine.

Can distributions prove interdependence of variables?

Guenter’s work can be serialized.

1. Apply this analysis to hard disk storage to ensure its accuracy, then apply to genomes.
2. Create a set of arbitrary definitions for sequence
3. Find those that are significantly common
   1. From the human genome, Venter ocean data, or transcriptome project
4. See if their distributions are non-poisson
5. Tabularize this data set and publisize it.

program: GA\_dnaorg.exe