Background

-Genomes have both useful genetic information coded in genes and presumed useless inter-gene regions termed as gaps.

-Gene pairs between organisms can be isolated using a tool called \textbf{B}asic \textbf{L}ocal \textbf{A}lignment \textbf{S}earch \textbf{T}ool, commonly referred to as BLAST.

- \textit{Salmonella Typhimurium LT2}, \textit{Citrobacter Koseri}, and \textit{Escherchia Coli MG1655} were chosen because they are closely related members of \textit{Enterobacteriaceae}.

- The phylogeny tree built from 16s rRNA shows the relative relations of the three bacteria chosen to be analyzed, which is an explanation of why E. Coli and Citrobacter have smaller common regions than Citrobacter and Salmonella.

**Motivation**

\item Mutation in genes has a double constrainment of DNA's built in error-handling as well as maintenance of gene function.

\item Gaps are suspected to have no function need a citation here probably, so they should mutate faster and more randomly.

\item choosing pairs of bacteria that were closely related could be used to show rates of mutation in homologous genes and gap regions.

\item Differing rates of mutation would suggest that gaps don't share gene's constrainment due to function

**Question**

\item Do gaps and genes mutate in the same way?

\item Is this trend consistent in different pairs of bacteria?

\item Do mutations show any form of dependence on other mutations?

**Hypothesis**

We hypothesize that genes will mutate at different rates than gaps because they have to code information essential to cell survival.

**Procedure**

Genomes were run through BLAST to identify matching regions for each pair comparison. BLAST matches were presumed to be homologous genes. Two matching regions with a non-matching region between them led the non-matching region to be labeled as a gap. Any gap longer than 500 bases was presumed to contain a gene that wasn't present in the compared genome and was therefore discarded.

Matched regions (both genes and gaps) were run through CLUSTALw and any base mismatch preceded and followed by at least 3 matching bases was presumed to be a mutation. Mutation rates were then tabulated, plotted, and chi-squared tested on mutation counts.

**Conclusions**

\item Purine->purine and pyramidine->pyramidine mutations still show as being markedly different, which shows uniformly random mutations aren't to be expected

\item

Genes and gaps have statistically different mutation rates as shown by a chi-square test (p $<$ \num{2.2} \* \num{10e-16}) in all three gene pairs

\item This suggests that genes and gaps have different mutation patterns

\item When insertions and deletions are removed, there's a large drop in statistical significance, which indicates that a major part of the difference is gaps losing information.

\item When insertions and deletions are removed, E Coli and Salmonella have the smallest change in significance, indicating that there are more base shift mutations in gaps than genes in that pair.

\item In the case of two mutated bases next to each other, there is Markovian dependence in genes

\item Two mutated bases following each other was rare enough in gaps to be unable to draw conclusions about dependence.