Shannon's uncertainty and Kullback-Leibler divergence in microbial genome and metagenome sequences

Sajia Akhter, Robert A Edwards

Computational Science Research Center, San Diego State University 2008

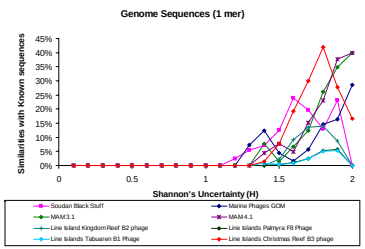
Shannon's Uncertainty on Genome Sequences

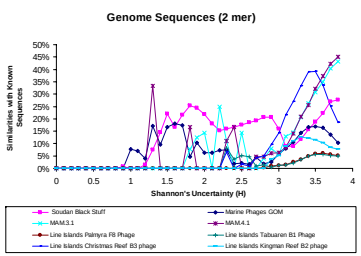
All genome sequence data contains inherent information in it. **Shannon's uncertainty** theory can be used to measure how much information a sequence has. Shannon's uncertainty can be calculated on DNA sequences by –

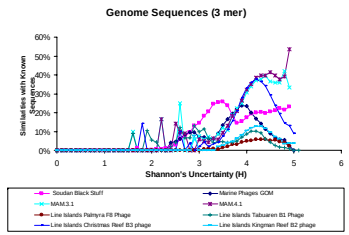
$$H = -\sum_{i=1}^{M} P(i)\log_2 P(i)$$

where P(i) is the probability of the occurrence of each base, M = 4 (A,G,C,T) for 1 mer, M = 16 (AC, AG, …) for 2 mer, and so on.

Here Shannon's uncertainty has been calculated for 24 datasets and it shows that the amount of information in a sequence correlates with the similar sequences that will be found in the database using search algorithms (BLAST).



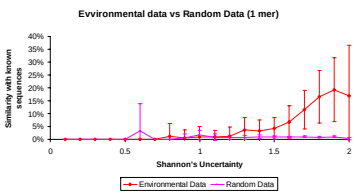




Hence, a sequence with more information (higher uncertainty), has a higher probability of being significantly similar to other sequences in the database.

Comparing with Random Sequences

For comparison, Shannon's uncertainty has been computed for 100,000 random sequences and also calculated the availability in the database using BLAST.



- For random sequences, the percentage of being similar with others is very small compared to the environmental sequences.
- Environmental sequences are much more likely to be similar to known sequences than a random set of sequences with the same uncertainty.

So, there is a strong correlation between Shannon's uncertainty and similarity among known sequences.

Kullback-Leibler Divergence on Amino Acid Composition of Bacterial Genome

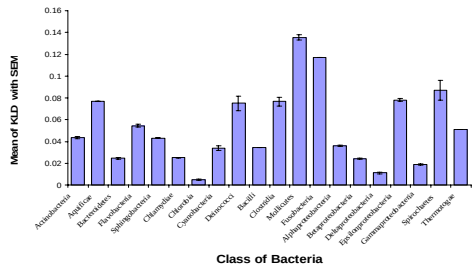
Kullback-Leibler Divergence is a way to measure the difference in amino acid composition of two samples. It calculates the difference between two probability distributions: from a "true" probability distribution *P* to an arbitrary probability distribution *Q*.

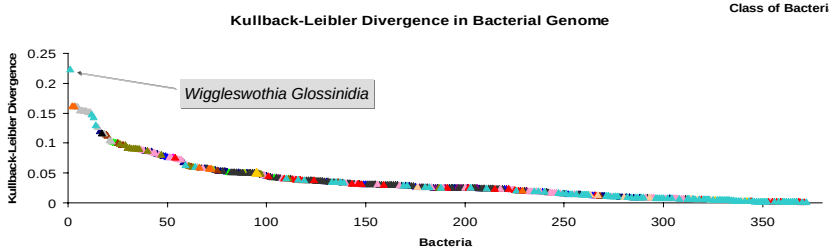
$$D_{KL}(P \parallel Q) = \sum_i P(i)\log \frac{P(i)}{Q(i)}$$

where P is the amino acid frequencies for each of the bacterial genomes and Q is the amino acid frequencies of all bacterial genomes. Here Kullback-Leibler Divergence has been calculated for 372 bacterial genome sequences.

The top ten genomes whose amino acid composition was most divergent from the mean were **endosymbionts**.

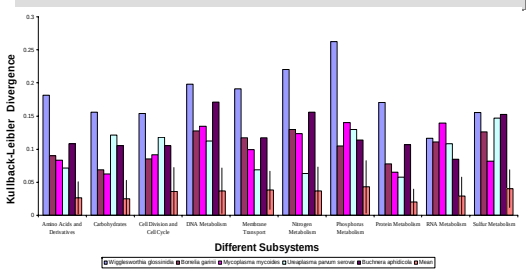
There is a significant difference between amino acid utilization in different phylogenetic groups of bacteria.





The first 5 bacteria having higher Kullback-Leibler Divergence –

- are significantly different from the mean for all subsystems.
- The differences are not restricted to one or few metabolic process but are across all subsystems.



Taking 5 random bacteria –

- there is not that much difference from the mean for all subsystems.