Proteins are the molecules that make life possible. They are composed of amino acid sequences with over 100 million on record. Protein sequences are random, by and large. This makes it challenging to infer component identity and functions, given partial information about a sequence. The project focuses on the uncertainty of inferring functions, based on the information provided by natural proteomes. Significantly, functional uncertainty was found to be a conserved property across archetypal proteins and proteomes. Further, the information of at least 1000 proteins is required for uncertainty for maximum correlations and conservation.

### Abstract

Information and uncertainty depend on our state of knowledge. In the case of protein, our knowledge depends on our access to proteome data. These serve as guides for spelling and grammar rules of protein sequences. The information is measured by the Shannon information. If we have a maximum uncertainty about a given component site in a protein, such corresponds to two bits because $2^{2} = 4$, the exponent quantifying the information in bits.

Lysozyme is a hydrolase that catalyzes the breakdown of bacterial cell walls. It is a vital tool of our immune system. The primary structure is represented by the amino acid component sequence, which underpins all the folding and functional properties. Let's take for instance a part of the lysozyme sequence (9) NWMC?AKWESG. The choices of our guesses are the fundamental functions of amino acids conferred by the genetic code: non-polar, polar, acidic, and basic. The non-polar amino acids are: A, V, L, I, P, F, W, M. The polar amino acids are: G, S, T, C, Y, N, Q. The acidic amino acids are D and E. The basic amino acids are K, R, and H.

### Methods and Results

#### Conclusion

The major takeaway from these results is the functional uncertainty being conserved across organisms and evolution. The correspondence of the proteomes is not dependent upon an individual protein sequence. The Shannon information corresponds to an entropy measure, with the protein functions conferred by evolution for various proteomes corresponding to points along an adiabatic curve. We instantly notice from a proteome of 100 sequences the inconsistencies amongst the Shannon information. It is not until proteomes of 1000 sequences or more that we start to see a linear correlation.

The graphs show the comparison of Shannon information for lysozyme functions using shark and human proteomes. The linearity attests to the equivalence of the proteomes. Sharks are a much older organism along the evolutionary tree than humans. The functional spelling and grammar rules however are conserved. If this was not true, then the points along the graphs would have been scattered all over the place. For each proteome we used the same number of primary structures $10^6$.

### References