Preparing people to lead extraordinary lives

“Protein Primary Structure Information and the Conservation of Functional Uncertainty”

Omar Zahra and Daniel Graham
Department of Chemistry and Biochemistry, Loyola University Chicago.

Preparing people to lead extraordinary lives

Abstract

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.

Introduction

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. Lets 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

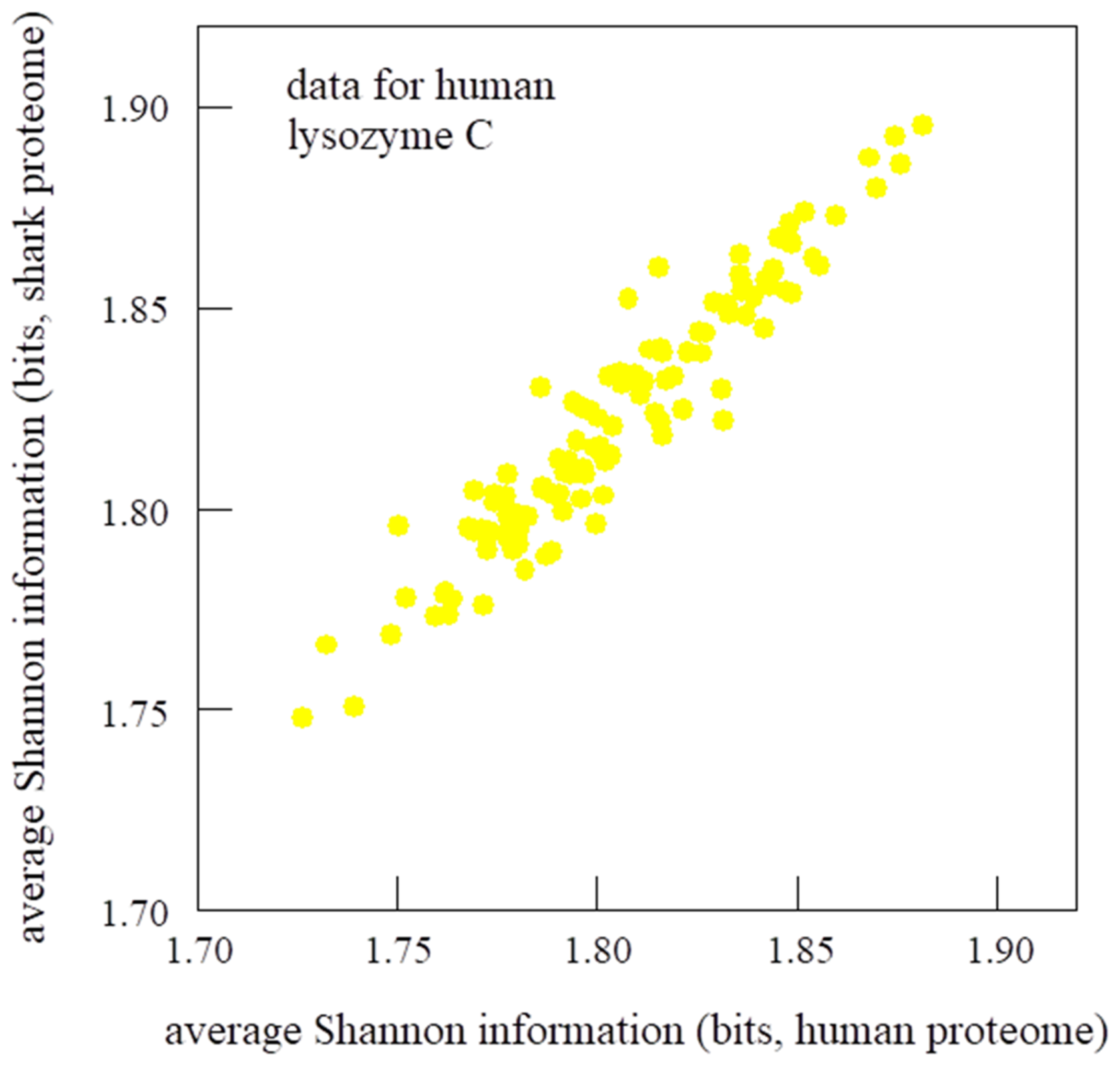

Figure 1

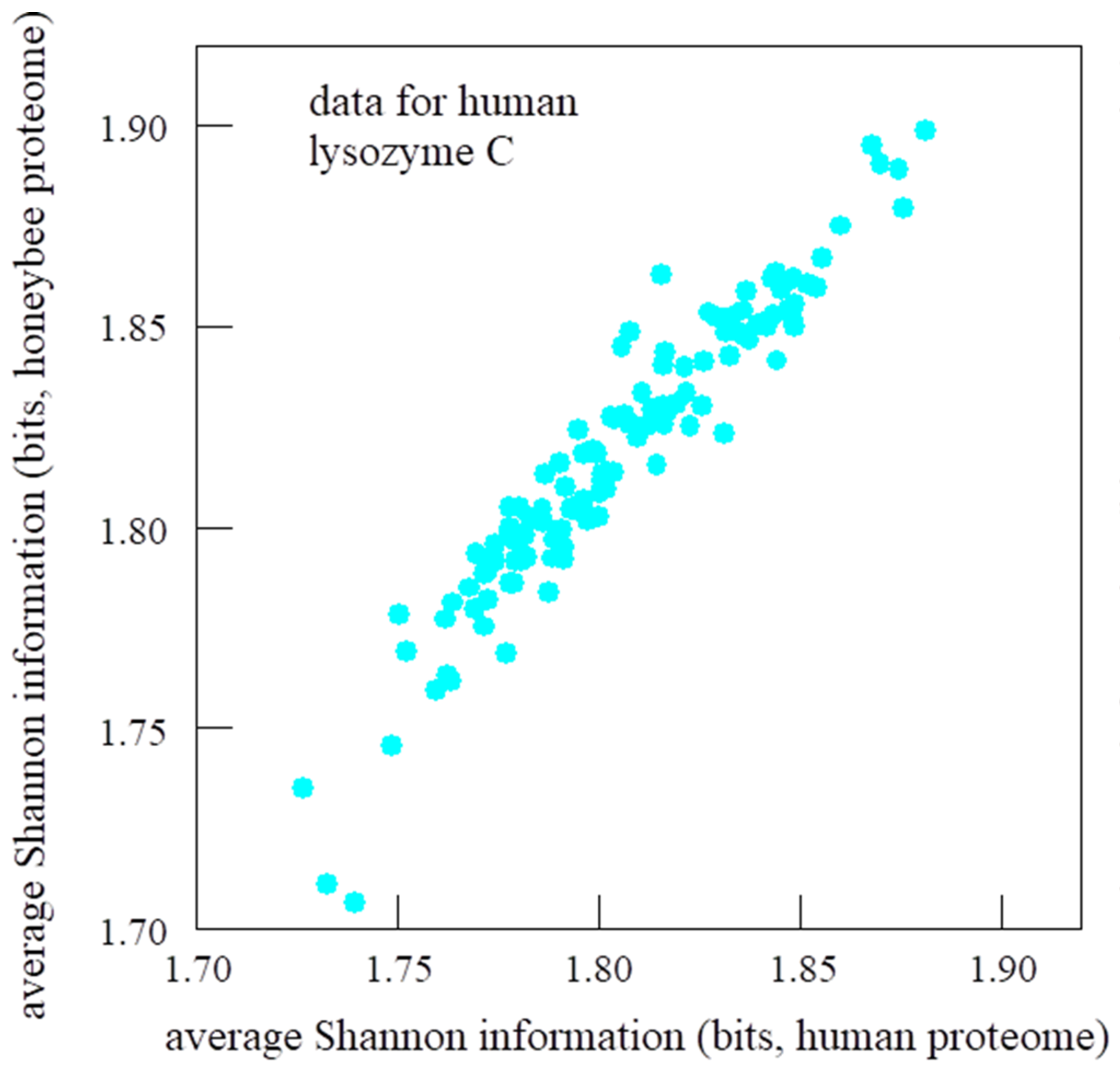

Figure 2

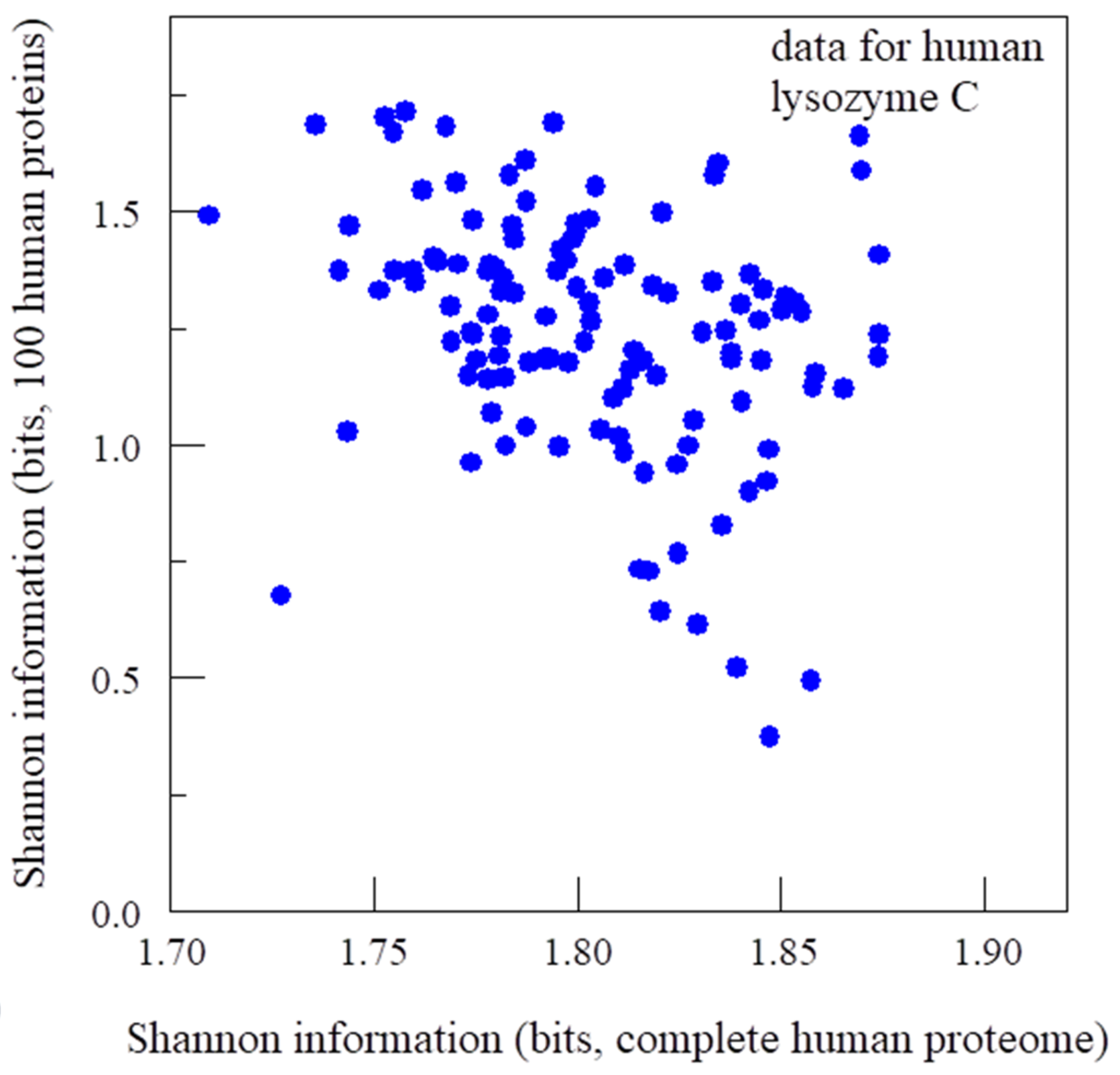

Figure 3

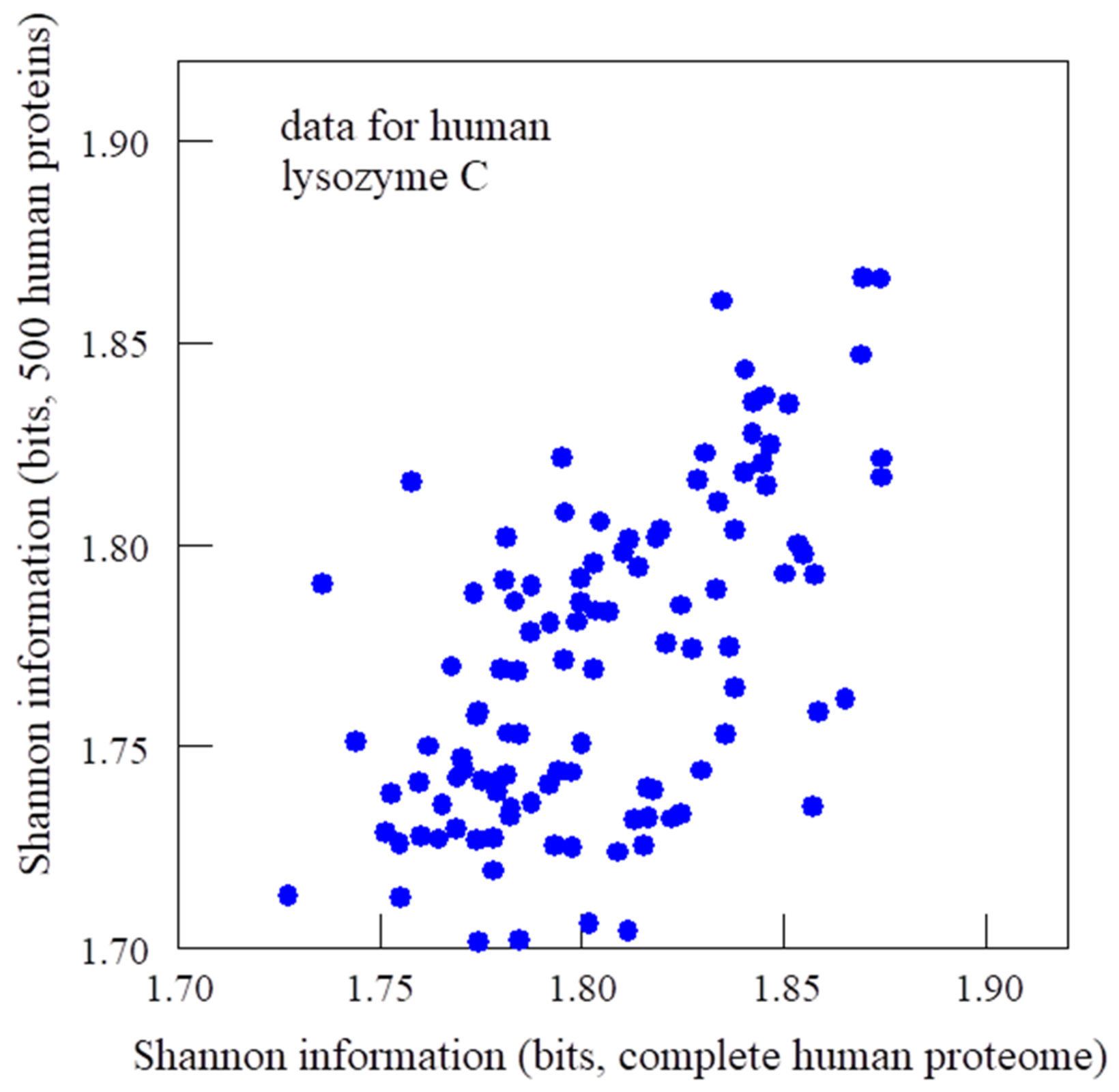

Figure 4

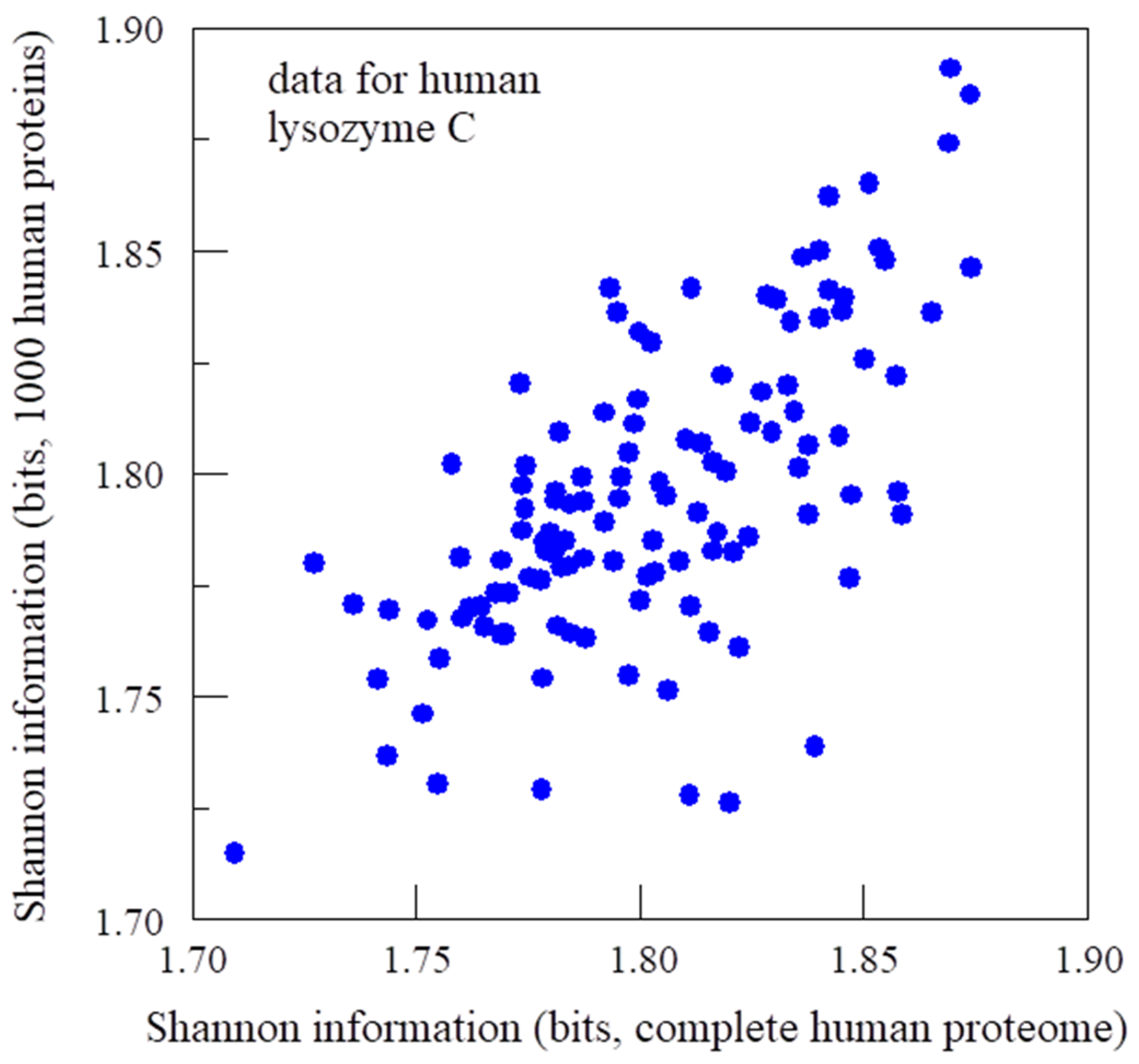

Figure 5

Figure 6

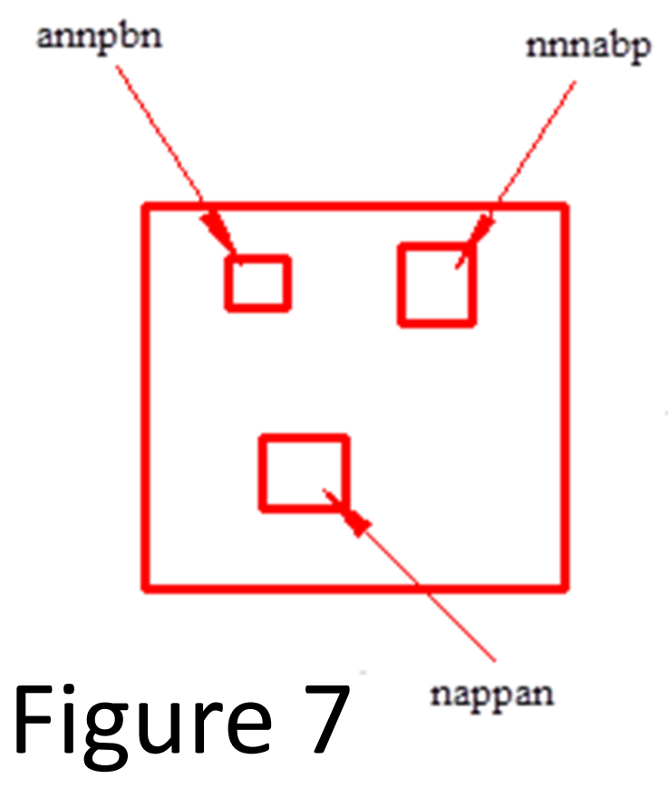

Figure 7

$$I = -\sum_{i=1}^{i=4} p_i \log_2 p_i$$

Figure 8 The units are in bits on account of using base-two logarithms.

KVFERCELARTLKRLGMDGYRGISLANWMC
LAKWESGYNTRATNYNAGDRSTDYGIFQINS
RYWCNDGKTPGAVNACHLSCSALLQDNIAD
AVACAKRVVRDPQGIRAWVAWRNRCQNR
DVRQYVQGCGV

Figure 9 Sequence for human lysozyme C

❑The choice of proteome is immaterial. The spelling and grammar rules are universal across organisms and evolutionary time scales. The uncertainty of chemical functions is independent of the choice of proteomes for guidance. Functional uncertainty for a protein is highly conserved

❑There is a significant size-dependence. We require at least 1000 proteins for nominal fluency. We acquire near-maximum fluency if we look to spelling and grammar guides of 5000 proteins or more. Proteomes of 100-500 proteins are nearly worthless.

We wish to

1. Quantify the functional uncertainty of sites in bits, making use of the Shannon formula and the primary structures of proteomes established by modern genomics.
2. Assess the dependence of uncertainty on the choice of proteomes. For example, using the proteomes of humans, sharks, honeybees, drosophila, rice, etc. to be the guides.
3. Assess the uncertainty dependence on the size of the chosen proteome. For example, using a complete or partial proteome as our grammar and spelling guides.

We scan the whole or partial proteomes for information across hexamer. Given four fundamental functions, this corresponds to a measure of space of size 4^6. Figure (6) represents the space. Scanning the data enables measured quantification of each hexamer possibility: annpbn, nnnabp, nappan, etc. Each possibility corresponds to a sector or parcel in the large space. The space is diverse in terms of parcel size. (Refer to Figure 8). Each protein of interest in respelled in terms of its fundamental functions. For example, lysozyme respells as:
bnnabpannbpnbbnpnappbpnpnnpnnpnnbnappppbnppppnpabppappnnpnpp
bpnppapbpnpnnpnpbnpppnnnpapnnannnpnbbnnbanppnbnnnnnbpbpppbanb
ppnppppn
Each hexamer in the protein presents as, for example, npnnpn. Then each site of the hexamer presents hints and uncertainty according to the neighbors, for example: ?pnnpn. We assess the probability of the ?-site being a given function, given the hints provided by the nearest neighbors pnnpn. There is a probability determined for each function. The probability obtains from looking parcel areas over the measure space for the proteome. The Shannon information formula then involves a sum over four terms (Figure 8).

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 attest to the equivalence of the proteomes. Sharks are a much older organisms 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 10^4.

References

- Roy, C., Wise, R., Jurukovski, V., DeSaix, J., Choi, J., & Avissar, Y. (2018, June 06). Genomics and Proteomics. Retrieved from https://bio.libretexts.org /Bookshelves/Introductory_and_General_Biology/Book:_General_Biology_(OpenStax)/3:_Genetics/17:_Biotechnology_and_Genomics/17.5:_Genomics_and_Proteomics
- Serrano, L. (2017, November 05). Shannon Entropy, Information Gain, and Picking Balls from Buckets (T. Man & H. Plata, Eds.). Retrieved from https://medium .com/udacity /shannon-entropy-information-gain-and-picking-balls-from-buckets-5810d35d54b4

LOYOLA
UNIVERSITY CHICAGO
Preparing people to lead extraordinary lives