1 Information Theoretic Constraints

What is the entropy of the human proteome, which we will call P? There are $N_P = 20139$ entries in the canonical Uniprot human proteome. If we assume a uniform distribution for protein frequency (i.e. an upper bound on the entropy) we find that

$H(P) = log_2(N_P) = 14.29$ bits / protein

This means that we need, on average, to extract a unique 15 bit binary word ("barcode") from a protein in order to unambiguously identify it.

If we construct our barcode by chemically labelling two amino acid residues, do those particular residues occur sufficently often and in a sufficiently random fashion to provide at least 15 bits of unique information? The total human proteome is composed of approximately 10^7 amino acids, and their frequency distribution can be measured.

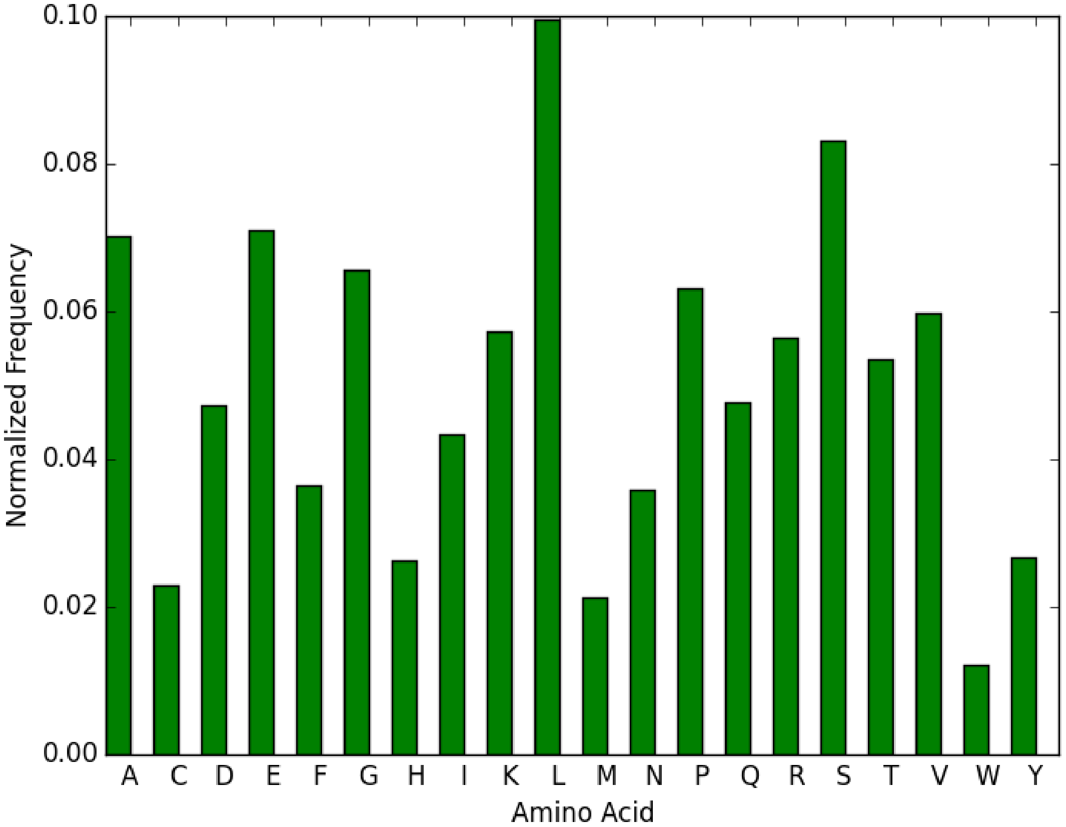

Figure 1:

The distribution of protein lengths can also be quantified.

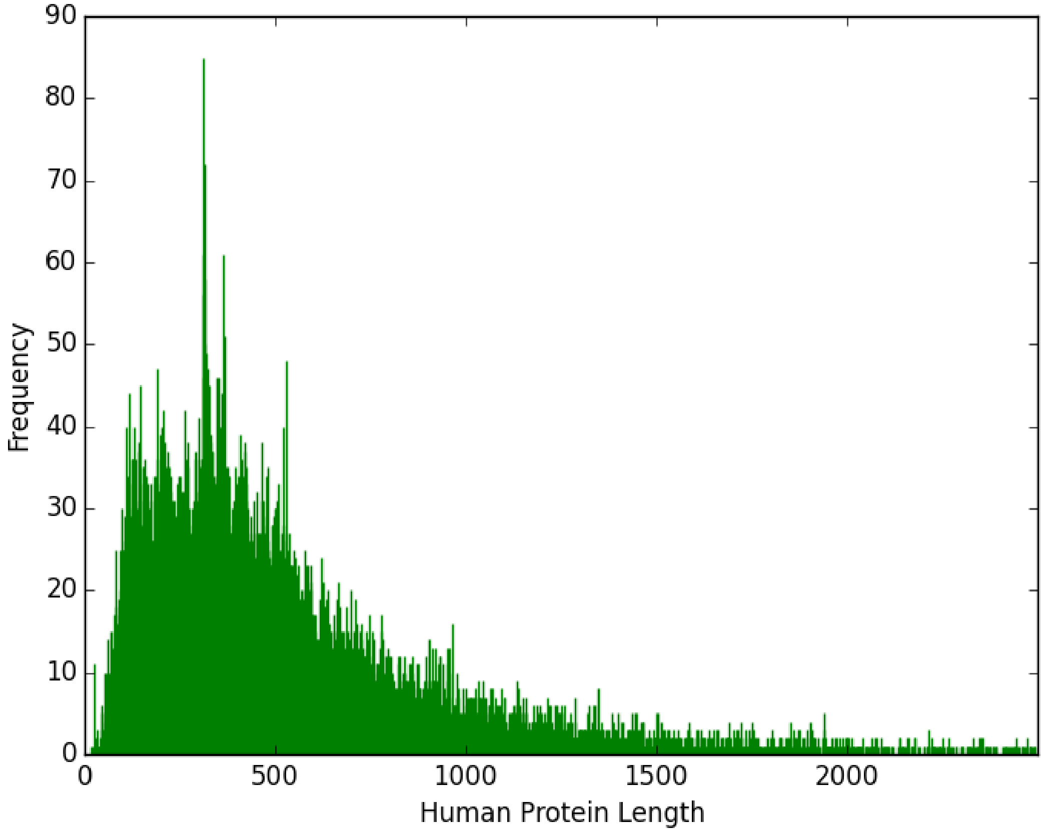

Figure 2:

This distribution yields an average protein length of $\langle l \rangle$ =561 amino acid residues. Combining this information with the information in 1 we can make rough predictions regarding the expected average size of a two-bit barcode. Taking the two most frequently occuring amino acids, L and S, occuring with frequencies f_L and f_S, the average number that occur per protein will be $\langle LS \rangle$ = $(f_L + f_S)\langle l \rangle$ or 102 L and S residues. Similarly, for the least frequent residues, M and W, we have $\langle MW \rangle$= 18. Therefore, to a very rough first approximation, we have between 18 and 102 bits to form our barcode with, and this is sufficent to meet the minimum barcoding requirement of 14 bits/protein. However, we have so far been considering average values only, and the real situation will be potentially complicated by factors incuding but not limited to:

- The median protein length being smaller than the average protein length due to the long tail of the distribution.
- Nonrandom distribution of amino acids in protein.
- Experimental complications including readout.