How big is the "average" protein?

Proteins are often referred to as the main workhorses of the cell. One of their intriguing features is the way in which individual molecules all operating simultaneously can produce effects that are seen at the scale of the entire planet. For example, Rubisco, the most abundant protein on Earth, performs one of the central reactions in photosynthesis, namely carbon fixation, and is responsible for extracting Gigatons of carbon from the atmosphere each year. This is ≈ 10 times more than all the carbon dioxide emissions of humanity from power plants, car tailpipes etc. Yet, all of this busy chemical fixation is carried out by individual molecules each of which is only 55 kDa in mass. At a more human scale, ATP synthase, the machine responsible for synthesizing the ATP molecules that power our bodies, in every human churns out each day a mass of ATP comparable to our body weight itself.

The size of proteins such as Rubisco and ATP syntase and many others can be measured both geometrically in terms of how much space they take up and in terms of their sequence size as determined by the number of amino acids that are strung together to make the protein. Given that the average amino acid has a molecular mass of 100 Da, this means that Rubisco, for example, has roughly 550 amino acids making up its polypeptide chain. The spatial extent of soluble proteins and their sequence size often exhibit an approximate scaling property where the volume scales linearly with sequence size and thus the radii or diameters tend to scale as the sequence size to the 1/3. A simple rule of thumb for thinking about typical soluble proteins like the Rubisco monomer is that they have 2-3 nm diameter as illustrated in Figure 1 which shows not only Rubisco, but many other important proteins that make cells tick.

To be concrete about the proteome-wide characteristics of protein size, we start with *E. coli*. As shown in Figure 2A, such proteins have a distribution of sizes, with a median protein size in *E. coli* around 280 aa. The determination of the distribution of protein sizes has been undertaken using both physical (such as 2D gels) and bioinformatic methods and each has its own strengths and shortcomings. For example, in the 2D gel methods culminating in data like shown in Figure 2, only a fraction of the full set of proteins are identified. Alternatively, the bioinformatic methods which involve sequence gazing can induce bias by not accounting for proteins with shorter sequences. (106444). One interesting feature of many of these proteins is the relative sizes of enzymes and the substrates they interact with. For example, in metabolic pathways, the substrates are metabolites which usually have a mass less than 500 Dalton while the corresponding enzymes are larger than 30,000 Da. In the glycolysis pathway, sugars are processed to extract both energy and building blocks for further biosynthesis. This pathway is characterized by a host of protein machines, all of which are much larger than their sugar substrates, with examples shown in Figure 1 where we see the relative size of glucose and the enzymes that interact with it.

Similar studies have been made for the human proteome. In humans, the median protein length is roughly 350 aa and the mean is 476 aa (Scherer, pp. 68). The entire

distribution of protein lengths in humans is shown in Figure 2B. These values are consonant with the simple rule of thumb which posits the "typical" protein to consist of ≈ 300 aa, though clearly there is a systematic trend for eukaryotic proteins to be larger, often with more recognizable protein domains per protein. Of course, one of the charms of biology is that it is chock-full of exceptions and when it comes to protein size, titin is a whopper of an exception. Titin is a multi-functional protein which behaves as a nonlinear spring in human muscles with its many domains unfolding and refolding in the presence of forces. Even though a typical protein is only 300 aa in size, titin is more than 100 times larger with its 33,423 aa peptide chain. Though it seems possible to identify the largest protein coded for in the genome, identifying the smallest proteins in the genome is still controversial (Skovgaard et al, 2001), but short ribosomal proteins of about 100 aa are common. An impression of the relative sizes of different proteins can be garnered from the gallery of different proteins shown in Fig. 1.

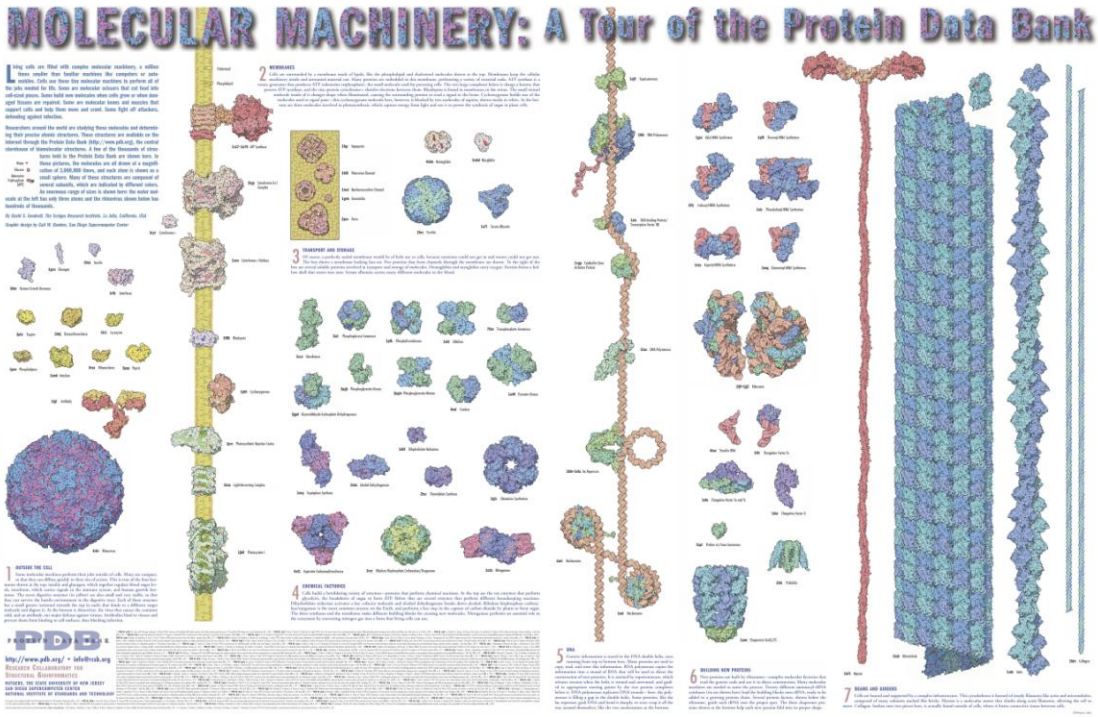

Figure 1: Gallery of proteins. Representative examples from some of the key functional roles of proteins are shown, all on the same scale to give an impression of the relative sizes of proteins. Need to include titin, rubisco, ATP synthase. Probably should comment on world's 2nd most important molecule.

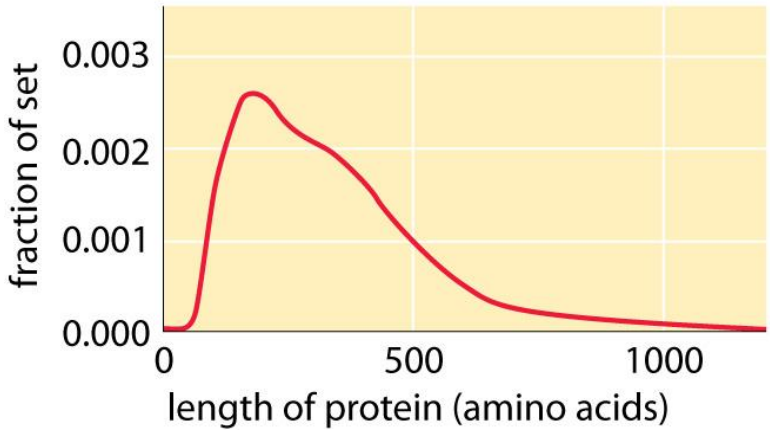

Figure 2: Distribution of protein lengths. (A) E. coli. (B) Human.

from On the total number of genes and their length distribution
in complete microbial genomes - Marie Skovgaard, Lars Juhl Jensen, Søren Brunak, David Ussery and Anders Krogh

Table 1: Median length of proteins in different species. The entries in this table are based upon a bioinformatic analysis of sequenced genomes.
(Brocchieri and Karlin 2005, 106444)

Organism	Median protein length (amino acids)
Homo sapiens	375[106445]
Drosophila melanogaster	373[106446]
Caenorhabditis elegans	344[106447]
Saccharomyces cerevisiae	379[106448]
Arabidopsis thaliana	356[106449]
5 Eukaryotes (above)	361[106450]
67 bacteria	267[106452]
15 archaea	247[106451]