Most Proteins Don't Exist!

Proteins fascinate me, they always have. Were it not for structural proteins, catalytic proteins (enzymes), transport proteins, binding proteins, etc., etc., there would be no conscious life. Arguably the only purpose of DNA is to encode the information required to build a protein. There are good arguments that the first self replicating life forms were probably RNA like nucleic acids, but life as we know it, life as we see it, is all about proteins. Perhaps in evolutionary terms, proteins are the ultimate parasite.

It is just a little over fifty years since Max Perutz and John Kendrew were awarded the Nobel Prize for determining the three dimensional structure of proteins, known as the tertiary structure. Four years before Perutz and Kendrew received their Nobel Prize, Fred Sanger had received his first Nobel Prize, for determining the amino acid sequence (primary structure) of a protein. Knowing the amino acid sequence of a protein and its three dimensional structure allows us to understand how it works. I remember an apocryphal tale in which Perutz predicted how long it would be before an undergraduate degree finals examination would include a question asking the student to design a protein. I can't remember the exact details of when this was expected to happen, but it is true to say that progress has been slower than Perutz had predicted and we are now only just beginning to see the emergence of designer proteins.

What we tend to forget, or perhaps not realise in the first place is that most proteins do not exist. This raises the possibility that only a tiny, tiny fraction of all possible protein structures are capable of having a useful function, at least under the ambient conditions of the planet Earth. How lucky we are that some useful proteins have come into existence. This disturbing fact, that most proteins do not exist, was first brought to my attention by one of my lecturers, Dr. Alan Yarwood, when I was an undergraduate at the University of Durham. The example given was cytochrome c, though the calculations below are my own.

Cytochrome c is a haem containing protein, an essential part of the electron transport chain. Two of its properties, its solubility and its small size made it a useful model for probing the evolutionary (phylogenetic) relationships between eukaryotic organisms. It is difficult to define an average protein size, and I would estimate that an average soluble enzyme is about 300 to 400 amino acids in length. At around 100 amino acids, cytochrome c is definitely at the small end of the spectrum.

Since terrestrial proteins are built from 20 amino acids, there are 20^{100} possible combinations of a cytochrome c sized molecule. If we take the average molecular weight of an amino acid residue in a protein as 110, then if we had 1 mole of each of these possible proteins we would have a mass of 1.394×10^{132} grams. If we had *only one molecule* of each of these combinations, then from Avogadro's number we can calculate that the total mass would be 2.315×10^{105} kg.

Estimates of the mass of the known universe do vary but they tend to be of the order of 3×10^{52} to 1.5×10^{53} kg. If just one molecule of every possible 100 amino acid protein sequence existed, the total would outweigh the mass of the universe. Quite clearly most proteins do not exist.

Tony Auffret
March 2013