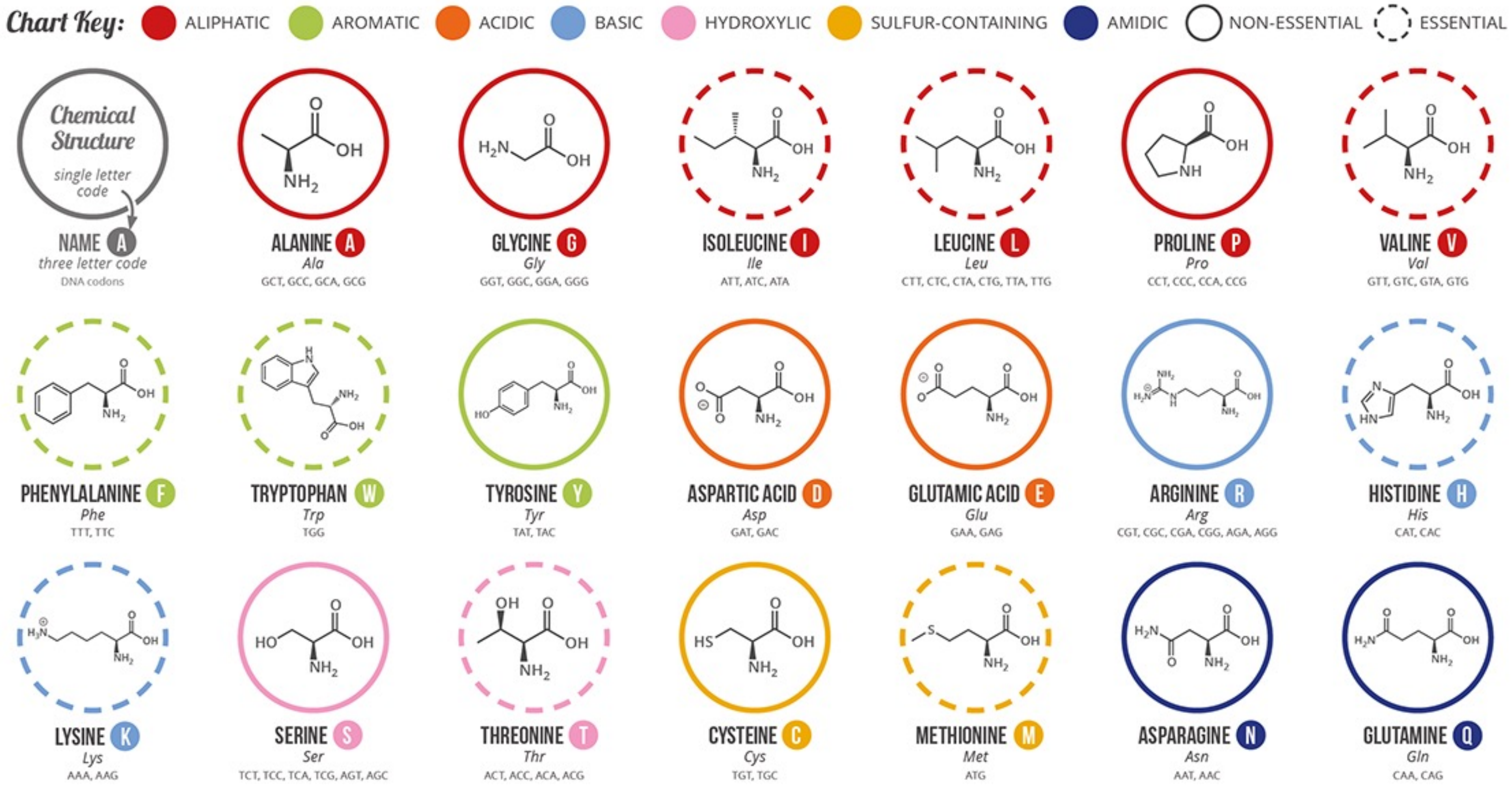
Homework 1

Amino acids and proteins

Amino acids (AAs) are biochemical compounds, which are the basic building blocks for proteins. Canonically, proteins in the human body use up to 20 different amino acids, which are designated with different letters of the alphabet.

Note that each of the AAs has also its unique chemical characteristics, which in turn give the resulting proteins unique properties.

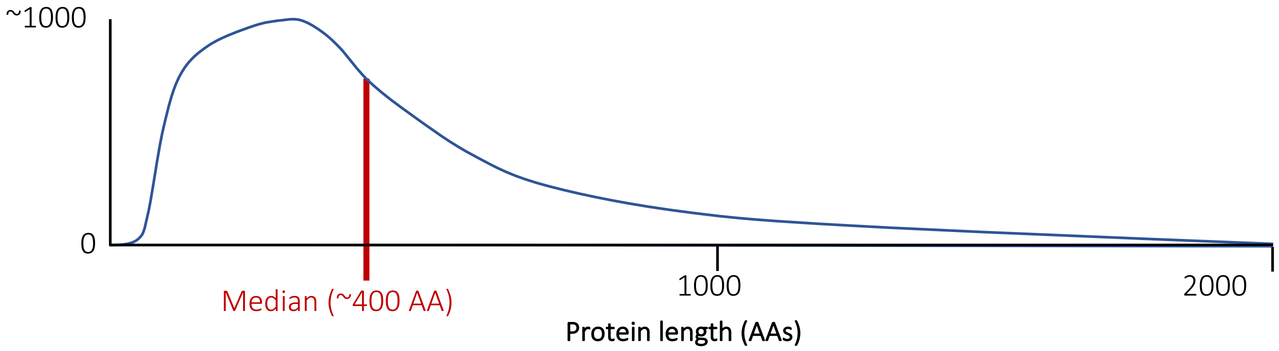
For the purposes of this course, it will **not** be necessary to know their structures or designations by heart.



Proteins

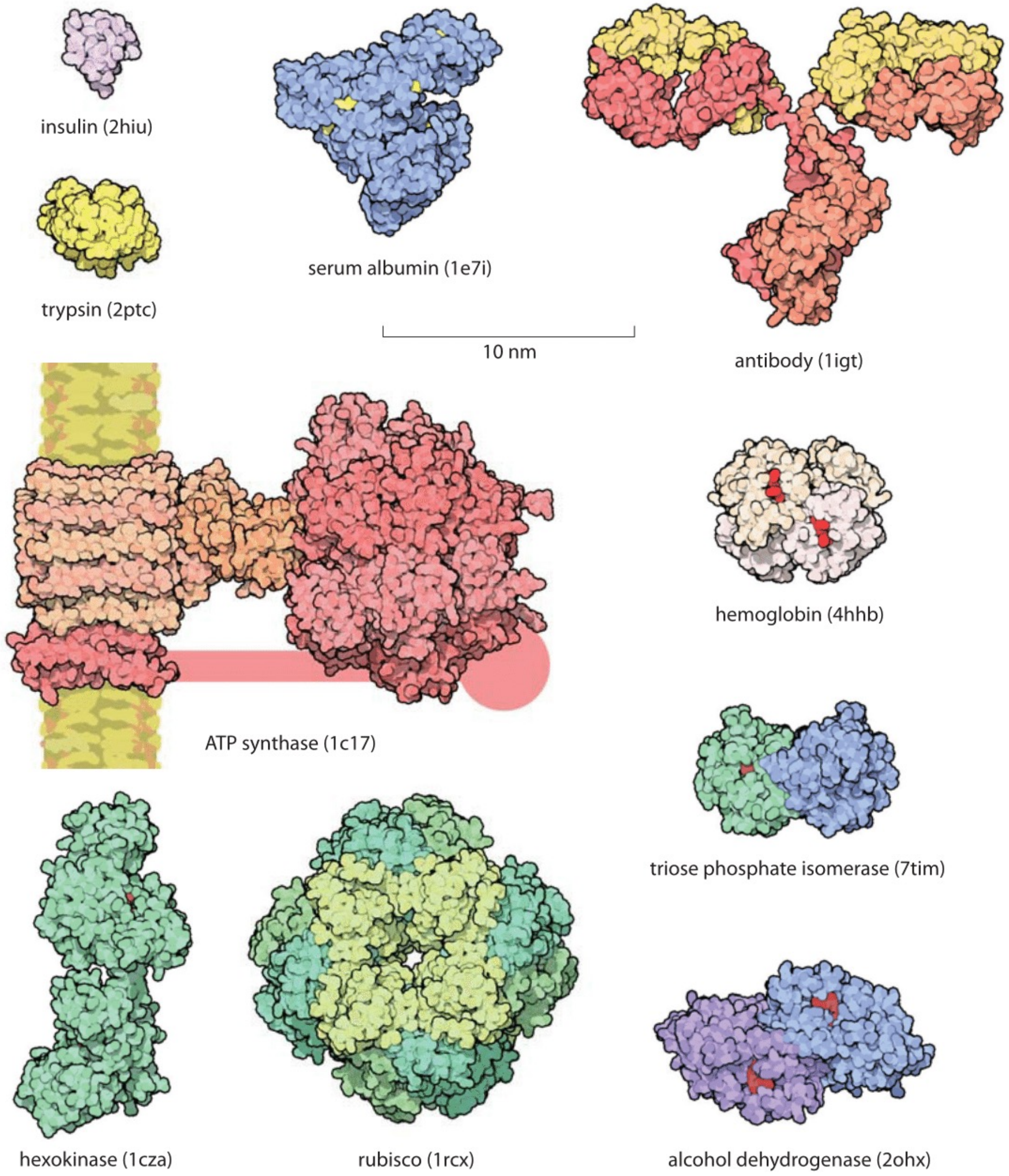
The human proteome (proteome = sum of all functional proteins) is estimated to consist of around 30,000 to 150,000 proteins.

The human proteome is considerably larger than the genome (genome = sum of all the functional genes), mostly due to alternative splicing.



Human proteome by length

On the right we see “spacefill” visualizations of different mammalian proteins, in their estimated three-dimensional conformations. Their unique “Protein Data Bank” code is listed in the brackets (https://www.rcsb.org/structure/2HIU).



http://book.bionumbers.org/how-big-is-the-average-protein/

FASTA format

“In bioinformatics and biochemistry, the FASTA format is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences. The format originates from the FASTA software package, but has now become a near universal standard in the field of bioinformatics.

The simplicity of FASTA format makes it easy to manipulate and parse sequences using text-processing tools and scripting languages like the R programming language, Python, Ruby, and Perl.”

For example, here’s the insulin protein sequence:

MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

(note that you can find the the amino acids and corresponding letters on the first slide)

Lets play around a bit…

You will each be given a unique human protein sequence.

1. Identify to which human gene your sequence belongs to:
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins (search might take a minute or two)
=> Which gene does your polypeptide sequence belong to?
=> Where is this protein located sub-cellularly? (search in https://www.uniprot.org/uniprot/ reports)
=> What might be the biological purpose of your gene/protein? Is it related to any diseases or biological conditions? (feel free to google around; https://www.genecards.org/ has short summaries; try the life sciences and biomedical research specific search engine https://pubmed.ncbi.nlm.nih.gov/; nowadays even Wikipedia will have decent information on many individual gene)

2. Find the gene itself:
https://genome-euro.ucsc.edu/cgi-bin/hgTracks?db=hg19
=> On which chromosome and where on it is your gene located?
=> Play around a bit – zoom in, change settings, click on elements…
=> Does your gene contain any conserved DNA sequences when compared to other species?
If yes, can you notice a pattern in terms of the conserved loci?
Why do you think that is?

3. In which tissue is your gene most highly expressed?
https://gtexportal.org/

Conserved sequence

Identical or highly similar sequences in DNA, RNA or proteins across species. Conservation indicates that a sequence has been maintained by natural selection.