Exercise 24: Which protein are you?

Andreas Loibl

November 17, 2010

At the Scansite Website of the MIT (`http://scansite.mit.edu/`) I searched for the expression "**ANDREAS**" and found one match:

Uncharacterized threonine-rich GPI-anchored glycoprotein PJ4664.02

(Protein-ID: YHU2_SCHPO, `http://www.uniprot.org/uniprot/Q96WV6.html`)

It's function is not known for sure, it is predicted to be a "cell surface glycoprotein", which means that it is located at the surface of the cell, anchored into a cell membrane by a glycophosphatidylinositol (GPI) anchor, and could therefore be used for inter-cellular-transportation or communication of signals into (or out of) the cell.

This protein was found in the "Schizosaccharomyces pombe" organism, which is a species of yeast (it is also called "fission yeast").

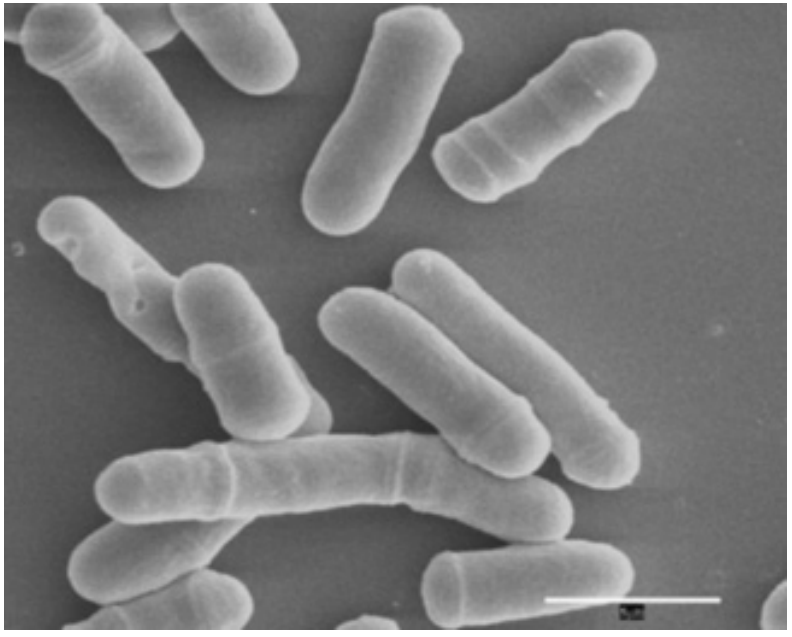

Figure: Microscopic view of the fission yeast. *[Source: Wikipedia]*

My name "ANDREAS" matches at the 101th- 107thamino acid, which is at 2.5% of the whole protein:

```
         10          20          30          40          50          60
MVEKTSPILI TLCIFLLIPI SSALPFQTFE ESLKLKSNIS HTGSFFANLS QSYSSTSIIH

         70          80          90         100         110         120
SVSSPVDCSL DVNNQINTIN ITSVSAGTNE LFLPTLSHAH ANDREASLFF PYTDIRYNGF

        130         140         150         160         170         180
PEPSNTDSTS ILSFNTIRLI PSTNVINTSH YKGFNRYPTI SSTVKVGKRA ATASFSYTNV

        190         200         210         220         230         240
SSSVIATAYT SASSTILSSP SVEQSTPSII TQTESSTTTE GSSVASSEST IANSDQSSFI

        250         260         270         280         290         300
TYESTQNPTA NKTDASQQST ESTSSSASAY SYITTLQTAT TAQQTTSENT YSTSGPNLTT

  (...)

        3910        3920        3930        3940        3950        3960
FNSTISLQPV VQFNNFTKRE ITTILITASD GSAVTTSLST FYSASSLASS VVKPLYILST

       3970
FFVAAVFFII F
```

Theoretically there would be a probability of

$$\frac{509019}{20^5} = 15.90\%$$

But I downloaded the whole Swiss-Prot database (`http://www.uniprot.org/downloads`) in FASTA format ($\sim$75MB gziped, $\sim$ 234MB extracted), extracted the starting sequence of all proteins and plotted a graph showing the frequencies of each amino acid

Probability of a given sequence of 5 amino-acids being the starting sequence of a protein

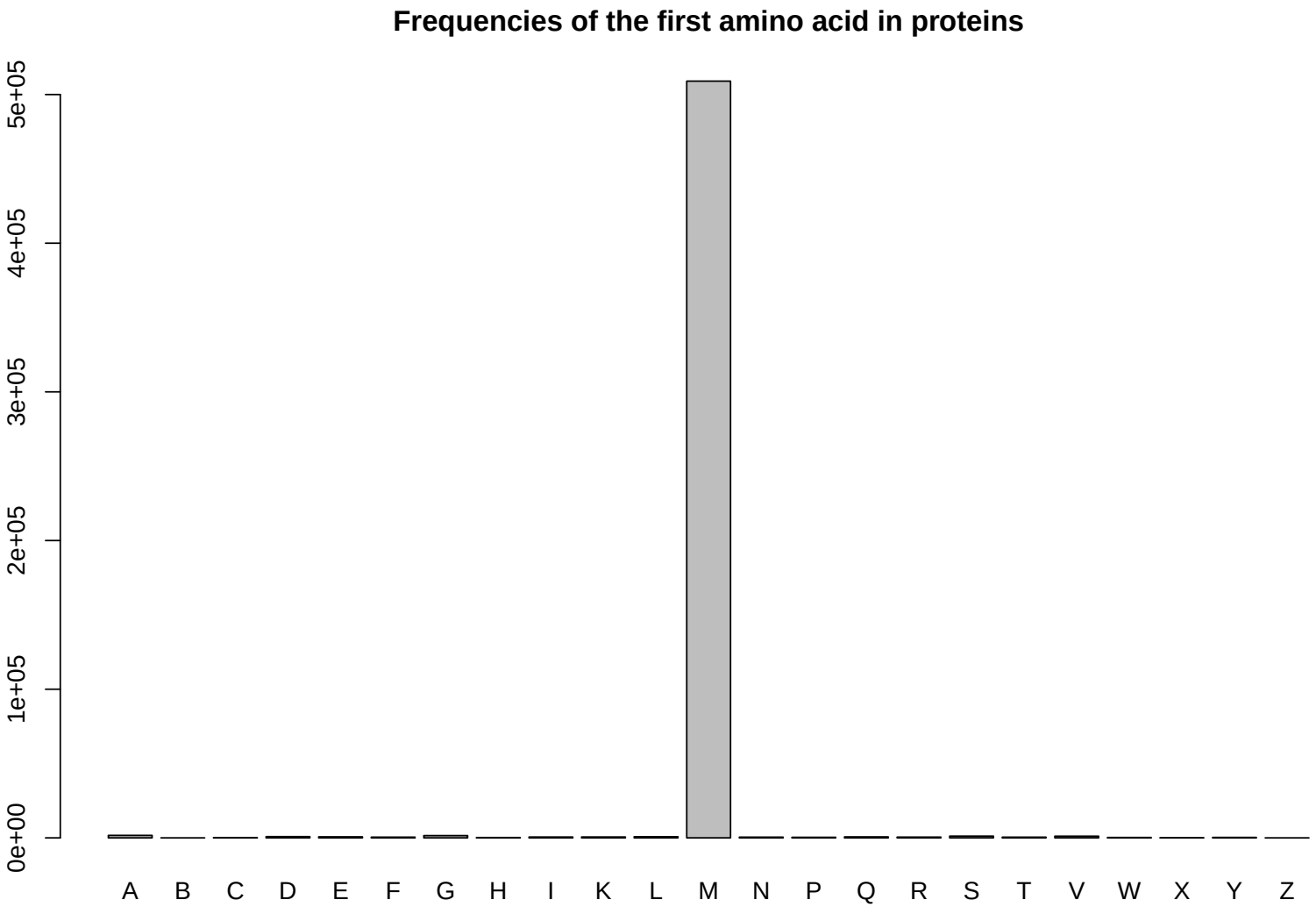

Figure: Frequencies of the first amino acid in proteins

So there is a over **97.5%** probability that the first amino acid is M (methionine). This can be explained by the fact that methionine encodes as "AUG", which is the start codon for almost all organisms. Let's have a look at the other amino acids:

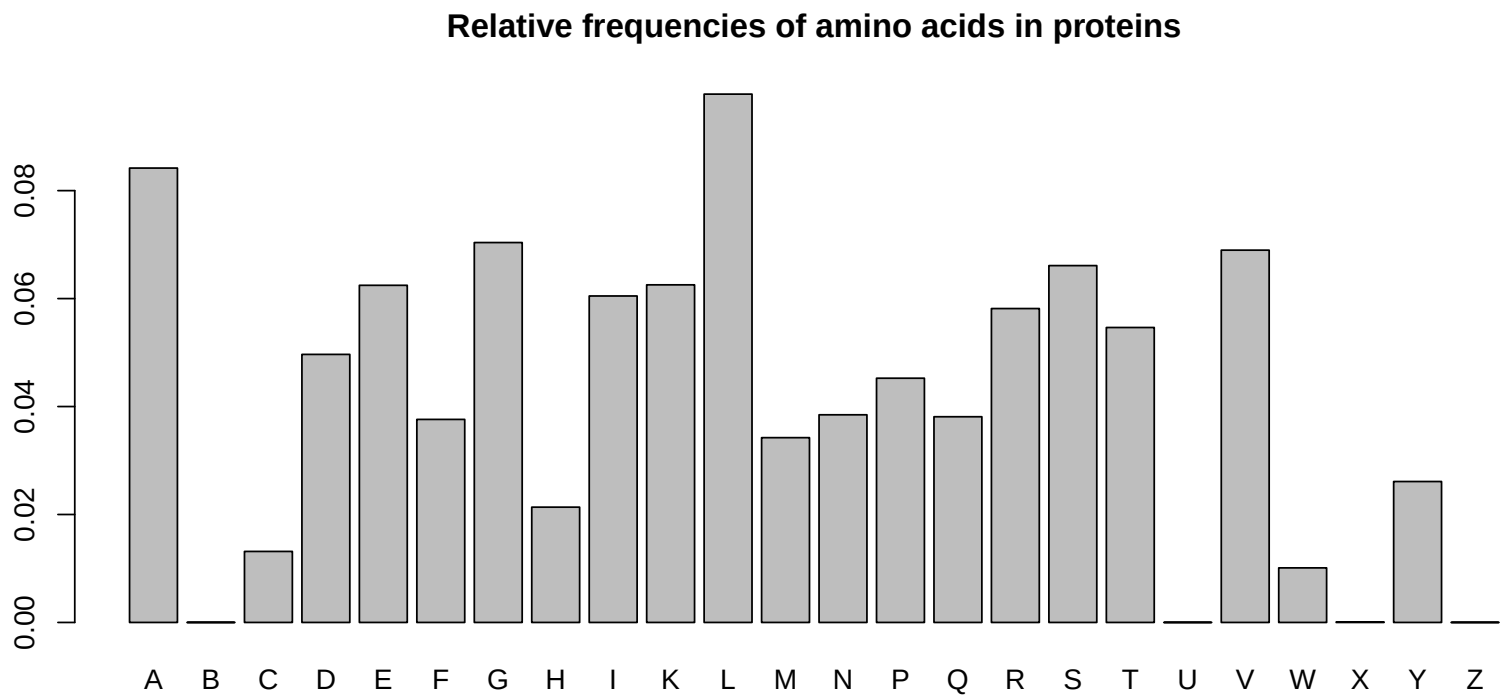

Figure: Relative frequencies of the amino acids in proteins

This shows that the second and the other amino acids are much more diversified, nevertheless they are not equally distributed, so you can't give an exact probability. But here at least an approximation:

$$p(a_1, a_2, a_3, a_4, a_5) = n \cdot \begin{cases} 97.5\% \cdot p(a_2) \cdot \ldots \cdot p(a_5) & \text{if } a_1 \text{ is "M"} \\ 0.13\% \cdot p(a_2) \cdot \ldots \cdot p(a_5) & \text{if } a_1 \text{ is not "M"} \end{cases}$$

$n = 509019$ is the number of all proteins and $p(a)$, the relative frequency of an amino acid, has to be taken from the graph in Figure 3 (page 8)

Examples: Probability of "MEGAN", "MANDY" and "PETER"

$$p(\text{"M"}, \text{"E"}, \text{"G"}, \text{"A"}, \text{"N"}) =$$

$$509019 \cdot 97.5\% \cdot 6.42\% \cdot 7.04\% \cdot 8.42\% \cdot 3.85\% \approx 740\%$$

$$p(\text{"M"}, \text{"A"}, \text{"N"}, \text{"D"}, \text{"Y"}) =$$

$$509019 \cdot 97.5\% \cdot 8.42\% \cdot 3.85\% \cdot 4.97\% \cdot 2.61\% \approx 210\%$$

$$p(\text{"P"}, \text{"E"}, \text{"T"}, \text{"E"}, \text{"R"}) =$$

$$509019 \cdot 0.13\% \cdot 6.42\% \cdot 5.46\% \cdot 6.42\% \cdot 5.82\% \approx 0.8\%$$

Note: a probability of over 100% means that there are probably more than one matches.