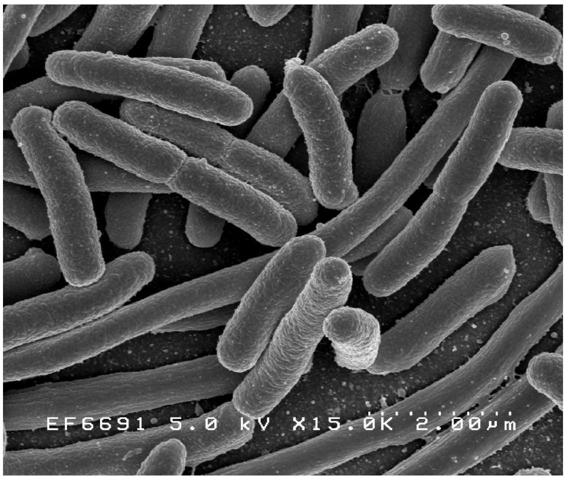
Instructions for the project.



Escherichia coli – picture: Rocky Mountain Laboratories, NIAID, NIH

The dataset selected for the project contains all predicted genes of bacterium Escherichia coli together with some of their calculated properties. Escherichia coli is one of the main species of bacteria that live in the lower intestines of mammals, known as gut flora. For most part the *E. coli* bacteria are harmless or beneficial – humans depend upon *E. coli* for the source of Vitamin K and B-complex vitamins. However, certain strains of E. coli (like, ***E. coli* O157:H7),** can be harmful and cause illness.

E. coli is widely used as a model organism for bacteria. Because of this long history of laboratory culture and manipulation, *E. coli* plays also an important role in modern biological engineering.

The dataset and its description are available on the course website: http://www.ms.ut.ee/BDA/m52orfs.txt (dataset);
http://www.ms.ut.ee/BDA/datadesc.pdf (detailed description of the dataset).

First tasks
Read the dataset description and import the dataset to R.

Second task
Look at the distribution of the variable pI. What is the most striking feature of the distribution?
Optional: Can you think of any possible explanation, why the distribution of variable pI has such a distinct shape?

Third task
Now look at the distribution of a) kD; b) ln(kD) (ln – natural logarithm, function `log` in R) . What is the main difference in the distribution of the variable in its original and logarithmic scale? Calculate the mean and median for both original and logarithmic values. In which of these two cases the median and the mean seem to be more similar?

Many classical statistical methods require the data to have approximate *normal* distribution. You can see, how a histogram of a perfectly normally distributed variable looks like by studying computer-generated random normal numbers, as done in the following example:

```
x = rnorm(1000, 10, 3)  # generate 1000 numbers from normal distribution
                        # with mean 10 and st.deviation 3
hist(x)
```

Which of the variables, kD or ln(kD) (if any) looks approximately normally distributed?

4th task

There exist some opinions, that the first or last amino-acid coding triplet of the gene can considerably influence the properties of the coded protein. Your task will be to investigate, whether the measured properties of the genes do depend from the starting codon or not. You may describe how the distributions of other variables vary from one starting or finishing codon to other. You may use descriptive statistics and graphics, whatever seems useful to you (no real statistical analysis needed yet). Please comment your results and highlight important discoveries.

How to present the project?

The project should be prepared using a word processor, e.g Word or OpenOffice Write. All numeric findings and graphs should be commented in the text. We recommend adding commented R code as an appendix in the end of the project.

Due to unexpected delay in sending out the project description, we have moved the deadline for submitting the project to December 28. The decisions on whose participation of the second part of the course will be financed will be mailed to you by January 3.