T-61.6070 Special course in bioinformatics I
Exercises 3.4.

1. **Biology.** The number of human genes is estimated to be around 20000 – 25000. The number of human proteins is estimated to be over 500000 (source: Wikipedia). How is this possible?

2. **Methods.** Explain briefly the hypothesis of co-evolving proteins (when considering protein-protein interactions).

3. **Probability model.** Calculate the conditional probabilities $P(D_i|D_1)$ $(i = 2a, 2b, 2c)$ with the equations below for the data in Table 1.

   - Which column i gives the highest probability and do you think this is plausible?
   - Can you explain why the model cannot distinguish between two of the columns? Do you think it should when a biological application is considered?

   (Note that you can consider columns 2a and 2b as different assingments where the first and second sequence (row) has swapped interaction partners.)

   Equations:

$$
\begin{aligned}
P(D_i|D_1) &= \frac{P(D_{1i})}{P(D_1)} \\
P(D_1) &= Z \prod_\alpha \frac{(n_\alpha^1 + 3)!}{3!} \\
P(D_{1i}) &= Z \prod_\alpha \prod_\beta (n_{\alpha\beta}^{1i})!
\end{aligned}
$$

   n_α^1 is the number of occurences of symbol α (A, T, G or C) in column 1. $n_{\alpha\beta}^{1i}$ is the number of occurences of symbol α in column 1 and symbol β in column i on the same row. "!" denotes the factorial $(0! = 1)$. Note that the normalization constant Z cancels out in the conditional probability and can be ignored.

Table 1: Data table.

1	2a	2b	2c
A	T	A	T
T	A	T	A
T	A	A	T
G	C	C	G
C	G	G	C