Nanopore Sequencing

Practical course M - Biological physics

Latest update: January 31, 2024

Imagine that you observe the change in behaviour or properties of an organism. This could be the acquisition of antibiotic resistance, a change in growth rate, or the sudden ability to move. How can you find out which change of the genome is responsible for this new behaviour? Nowadays, whole genome sequencing has become an important tool to address this type of questions.

Genome dynamics: mutation, insertion, deletion, and horizontal gene transfer

The genome of an organism comprises its hereditary information. This information is stored on the chromosome(s), DNA molecules with specific sequences. The DNA sequence is a linear succession of nucleotides each of which contains one out of the four bases adenine (A), cytosine (C), guanine (G), and thymine (T). For simplicity, we focus on bacteria which reproduce asexually and this means that the mother cell passes a copy of its chromosome to both daughter cells during reproduction. However, by various mechanisms, the DNA sequence of an organism changes with time (Fig. 1). During DNA replication or upon DNA damage, one base can be inserted, deleted or substituted by another base (Fig. 1**A**). These processes are called point mutations. Sometimes, a sequence is duplicated (Fig. 1**A**). Next to point mutations and duplications, horizontal gene transfer (HGT) heavily affects bacterial genome dynamics. During HGT, DNA from a different strain or species enters the cell and recombines with the chromosome. Transformation is one of the most common mechanisms of HGT (Fig. 2, [1], [2]). During transformation, competent cells take up free DNA from the environment (Fig. 2). Once taken up, recombination proteins (including

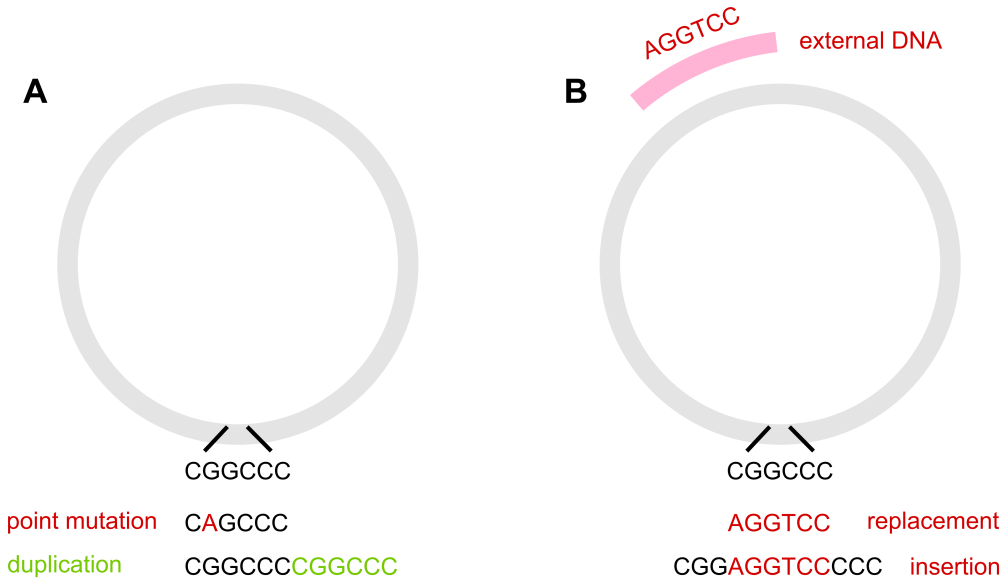

Figure 1: Genome dynamics. The chromosome is depicted in grey. **A**) In the absence of horizontal gene transfer, point mutations, deletions, and duplications usually dominate. Point mutation: one base (here guanine, G) is replaced by a different base (here adenine, A). Duplication: a segment of DNA is duplicated. **B**) Genome dynamics in the presence of horizontal gene transfer (Fig. 1). An external segment of DNA recombines with the chromosome. Here, the segment has a similar sequence compared to the sequence on the chromosome. It can therefore replace a segment of DNA appearing as multiple point mutations. The external DNA can also be inserted into the chromosome causing an increase in length.

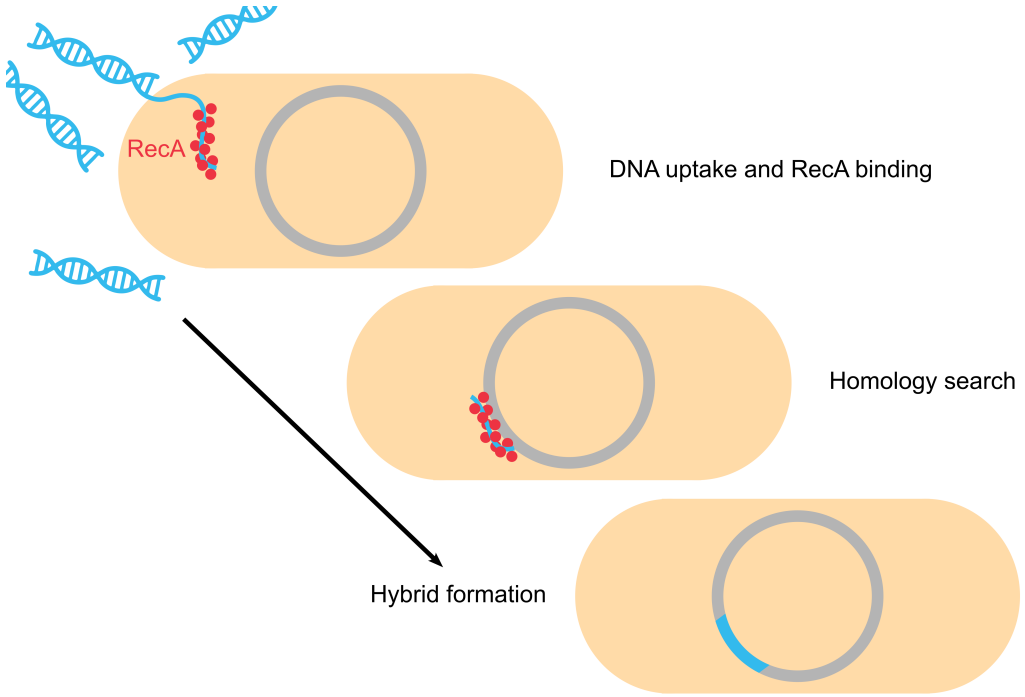

Figure 2: Horizontal gene transfer. Transformation is the uptake and integration of free DNA from the environment. During transformation, cells take up free DNA as single-stranded DNA. Once taken up, the DNA binds to the protein RecA which performs a homology search along the bacterial chromosome. When a similar sequence is found, a fragment of chromosomal DNA is replaced by the external DNA via homologous recombination.

RecA) bind the incoming single-stranded DNA and perform a homology search along the bacterial chromosome. When a similar sequence is found, the imported DNA can replace a segment of the chromosomal DNA via homologous recombination. This replacement is permanent and heritable. We would detect such a replacement event as multiple point mutations in very close proximity on the genome (Fig. 1**B**, [3]). External DNA can also be inserted into the chromosome, extending the entire length of the chromosome. Horizontal gene transfer can have important consequences on bacterial fitness, because genes and their associated functions can be acquired de novo. Often, bacteria acquire antibiotic resistance or virulence traits through horizontal gene transfer.

DNA sequencing

Mutations and horizontal gene transfer events are stochastic and occur at low rates. For example, the point mutation rate of bacteria is $\sim 10^{-8} - 10^{-9}$ nucleotide^{-1} generation^{-1}. Most mutations and gene transfer events decrease the fitness of bacteria, yet some enhance the fitness. Importantly, when bacteria with different mutations compete for growth resources, the ones with beneficial mutations are selected for. If we ask the question how bacteria adapt to a specific environment, then we need to know which genes carry mutations, which genes are deleted or newly acquired, etc. To find out how the DNA sequence has changed, we use different sequencing techniques, including Sanger sequencing, next generation sequencing, and nanopore sequencing:

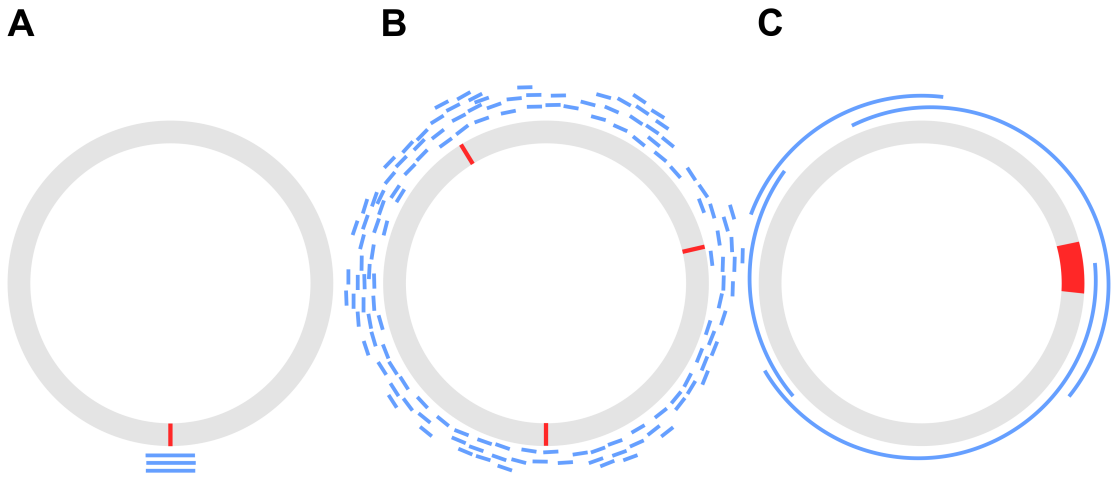

Figure 3: Various sequencing techniques. The chromosome is depicted in grey and sequence changes in red. **A**) Sanger sequencing yields the sequence of a specific fragment of DNA with a length of ≅ 1000 bp. **B**) Next generation sequencing provides the sequences of short (≈ 150 bp) overlapping segments across the entire genome. By mapping them to a similar known sequence, point mutations and replacements can be detected. **C**) Nanopore sequencing yield the sequences of segments with a length exceeding 1000 bp, sometimes up to 10^6 bp. Additional to the mutations, structural variations can be detected.

.

In case we know in which gene we expect to find a mutation, we use Sanger sequencing (Fig. 3**A**). This method provides the sequence of a specific segment of DNA that is no longer than ~1000 bp. If we don't know where on the genome mutations have occurred, then we have to sequence the whole genome. This is often done by means of next generation sequencing. This technique provides the sequences of overlapping DNA segments (reads) with a length of ~ 150 bp (Fig. 3**B**). The coverage of overlapping segments is typically in the range of several 100. The exact methods of Sanger and next generation sequencing are not subject to this course, but for further reading, please refer to [4] and [5]. Next generation sequencing works really well if the sequence of a very similar genome is already known. For example, the sequences of many strains of *Bacillus subtilis* are available on data bases [6]. If we were to perform a laboratory evolution experiment, then we would expect only few changes on the genome during adaptation to a novel environment, and the next generation sequencing reads can be mapped onto the reference genome. However, if we wanted to assemble a genome de novo, longer reads are required (Fig. 3**C**). Furthermore, if there are duplications or de novo insertions, then it is often impossible to localize them using the short reads. Nanopore sequencing is a novel technique that provides long reads. Sometimes an entire bacterial genome (several 10^6 bp) is obtained in one read.

Detection of single DNA molecules during their passage through nanopores

Membranes are the natural enclosure of each of our cells. They consist of lipid molecules that self-assemble into lipid bilayers (aka lipid membranes). In vitro, we can let lipids self-

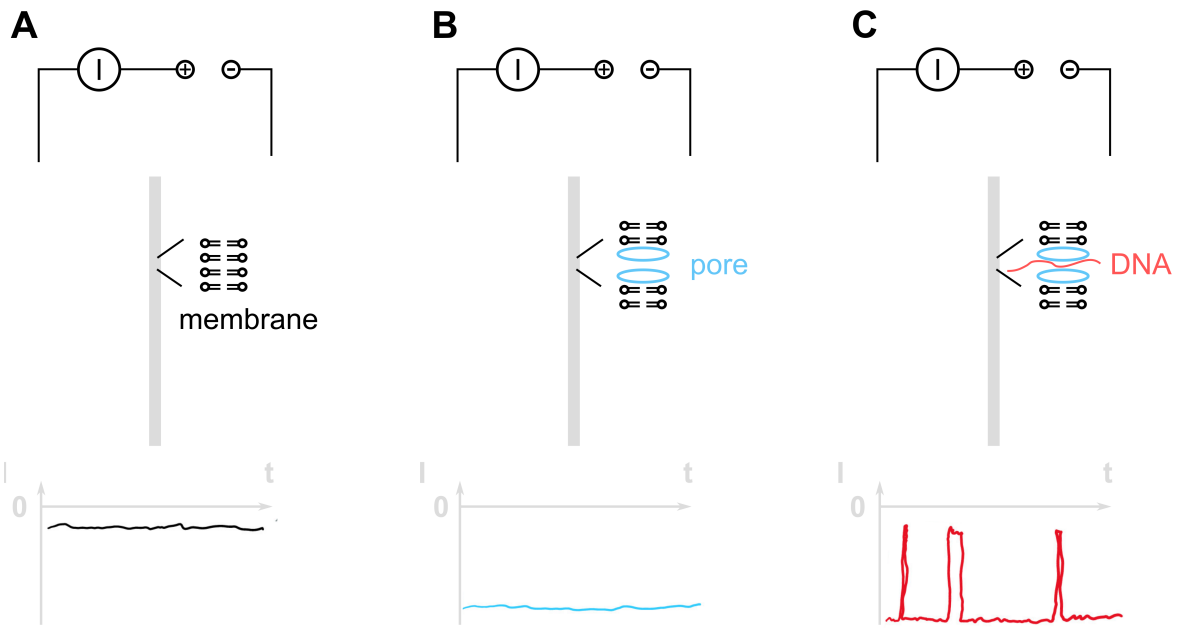

Figure 4: Setup for detecting single DNA molecules. **A**) An electrical voltage is applied across a membrane (grey) formed by lipid molecules (black) and the current is detected. **B**) Pore-forming proteins are inserted into the membrane. They decrease the electrical resistance. **C**) As single DNA molecules pass the pores, the increase in resistance is detected as a reduction of current.

.

assemble into membranes by inserting lipid molecules into water (Fig. 4**A**). When we apply an electrical voltage across the membrane, we find that the permeability to ions is very low resulting is very high electrical resistance. Bacteria generate pores or channels within their membrane to control the flux of ions and other molecules. Some of these pores (formed by proteins) can be reconstructed into lipid membranes (Fig. 4**B**). The electrical resistance is then governed by the pores. In the late 1990ies, Kasianowicz et al had the idea to use this synthetic system for detecting the length of single DNA molecules [7]. They inserted a pore called α-hemolysin into the membrane and showed that DNA moves through the channels in the presence of an electrical field. The crossing of DNA was detected as a transient increase in electrical resistance (Fig. 4). They found that the lifetimes of resistance increase was proportional to the lengths of the DNA molecules.

The experimental setup nanopore sequencing

Kasianowicz et al. foresaw that after several improvements their device could be used for sequencing. Each of the bases of a DNA molecule has a different chemical structure, and it was conceivable that the change of electrical resistance was different for each base. However, almost two decades of intensive research and development were necessary to bring this idea to the market. The first breakthrough idea was to use an enzyme called DNA polymerase instead of the electric field to drive DNA translocation through the pore [8]. This way that rate of translocation was slower and better controlled. Moreover, the pore-forming protein complex MspA replaced α-hemolysin providing a shorter contriction (Fig. 5**A**, [9]) to ensure that a lower number of bases contributes to the resistance increase.

The second major problem was that the resistance is affected not only by a single base re-

siding within the constriction, but also by the adjacent bases. This problem is illustrated in figure 5**B**. Manroa et al. measured the current through an MspA pore for a DNA template that contains only "CTA" nucleotide triplets with the exception of one "CGA" in the middle. When extracting the mean current levels (lower graph), a repeating pattern can be identified. The G substitution does not only affect one but four levels (changes highlighted in orange) and all to varying degrees. With this complication and the noise of the current data (upper graph) in mind, you get a feeling for how error-prone the conversion from signal to bases ("base-calling") is. To solve this problem, machine learning algorithms are applied with great success.

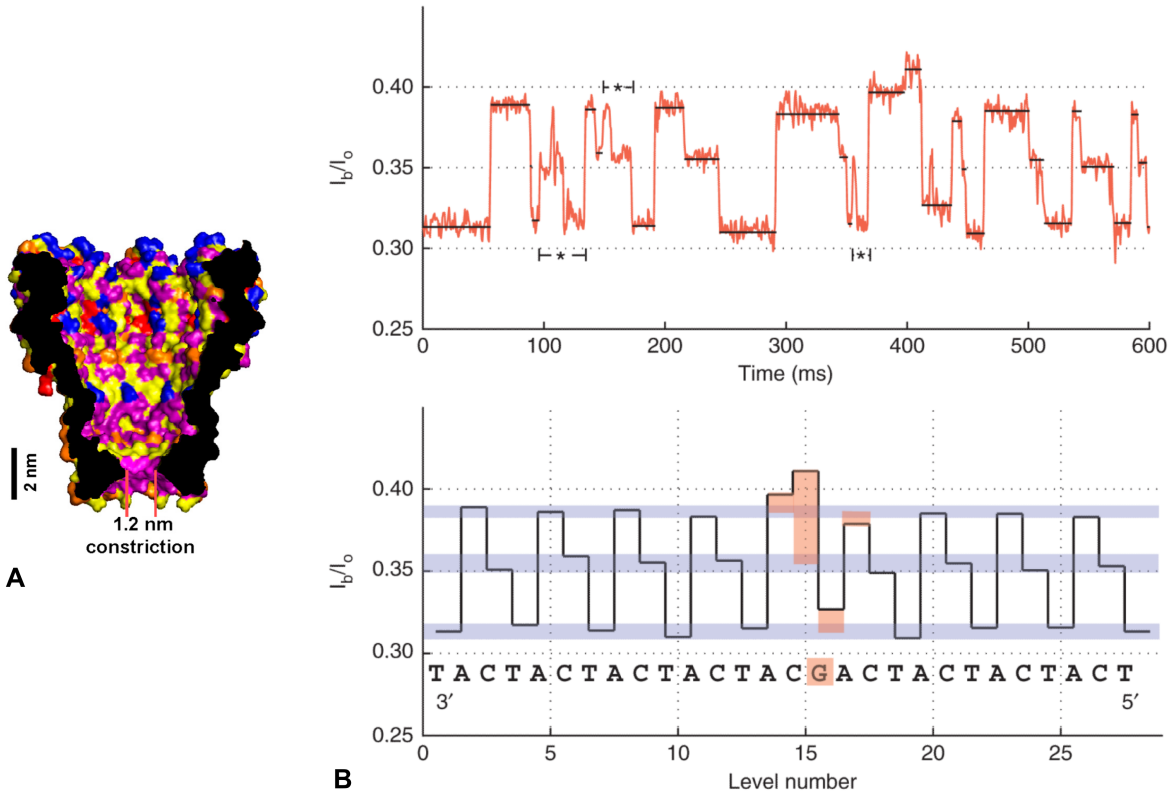

Figure 5: **A**) In a mutated form of the protein pore *Mycobacterium smegmatis* porin A (MspA) with a narrow and short contriction (1.2 nm wide, 0.6 nm long) a smaller number of nucleotides affects the current through the pore. Adapted from [9]. **B**) The upper graph shows the current measured for a DNA template that contains only "CTA" nucleotide triplets but one "CGA" in the middle (sequence shown in the lower graph, [8]). Extracting the mean current levels (lower graph) reveals a repeating pattern. The G substitution affects four levels, all to varying degrees (highlighted in orange).

Base-calling via machine learning

Base-calling is a big challenge in nanopore sequencing as shown in the previous chapter. Machine learning methods have proven highly useful for translating the current signal into a DNA sequence. This chapter aims at providing a rough idea of how deep learning methods are applied to sequencing. We note that the accuracy of base callers improves continously (Fig. 7), and state-of-the-art algorithms are too complex to be discussed here. Instead, we illustrate the method at the example of a base caller that is based on recurrent neural networks [10]. Recurrent neural networks (RNN) have been already used successfully for speech recognition [11], sleep phase modelling [12] and other sequence processing tasks.

From the input vector $\vec{x}_i$ which in this case consists of the mean, standard deviation, and length of each event i (Fig. 5**B**), the RNN calculates the output vector $\vec{y}_i$ that provides the probability distribution of called bases. The input is fed forward to multiple hidden layers while applying a learning rule. Each hidden state $\vec{h}_i$ depends on the previous hidden state $\vec{h}_{i-1}$ and the input: $\vec{h}_i = f(\vec{h}_{i-1}, \vec{x}_i)$. The functions f and g will be discussed later. This flow of information in forward direction is called feedforward propagation [13]. Increasing the number of layers can improve the precision. For three layers (as used bei Boža et al.) the calculations are as follows:

$$\vec{h}_i^{(1)} = f_1(\vec{h}_{i-1}^{(1)}, \vec{x}_i)$$
$$\vec{h}_i^{(2)} = f_2(\vec{h}_{i-1}^{(2)}, \vec{h}_i^{(1)})$$
$$\vec{h}_i^{(3)} = f_3(\vec{h}_{i-1}^{(3)}, \vec{h}_i^{(2)})$$
$$\vec{y}_i = g(\vec{h}_i^{(3)})$$

For example, the second hidden layer $\vec{h}_i^{(2)}$ depends on the previous second hidden layer $\vec{h}_{i-1}^{(2)}$ and the current first hidden layer $\vec{h}_i^{(1)}$ which depends on the previous first hidden layer $\vec{h}_{i-1}^{(1)}$ and the input $\vec{x}_i$ ... The connection between the hidden layers of different events accounts for the fact that the current signals are affected not only by the base that resides within the constriction, but also on previous bases. Note that f_1, f_2, f_3 are different functions with different sets of parameters. If the output $\vec{y}_i$ depends not only on the previous data points but also on the following points bi-directional RNNs are used (Fig. 6). In the bi-directional RNN the hidden layers depend on the hidden layers in both directions which looks as follows when using two hidden layers:

$$\vec{h}_i^{(1+)} = f_{1+}(\vec{h}_{i-1}^{(1+)}, \vec{x}_i)$$
$$\vec{h}_i^{(1-)} = f_{1-}(\vec{h}_{i+1}^{(1-)}, \vec{x}_i)$$
$$\vec{h}_i^{(1)} = \vec{h}_i^{(1+)} \| \vec{h}_i^{(1-)}$$
$$\vec{h}_i^{(2+)} = f_{2+}(\vec{h}_{i-1}^{(2+)}, \vec{h}_i^{(1+)})$$
$$\vec{h}_i^{(2-)} = f_{2-}(\vec{h}_{i+1}^{(2-)}, \vec{h}_i^{(1-)})$$
$$\vec{h}_i^{(2)} = \vec{h}_i^{(2+)} \| \vec{h}_i^{(2-)}$$
$$\vec{y}_i = g(\vec{h}_i^{(2)})$$

where $\|$ is the concatenation of vectors. The three-hidden-layers case is shown in figure 6 visualizing the underlying connections. f and g are called activation functions. They help the neural network to use important information while suppressing less relevant ones. Activation functions need to be non-linear. Otherwise all layers of the neural network will

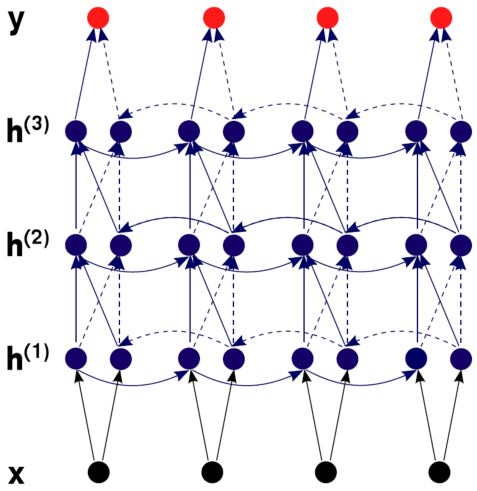

Figure 6: Schematics of a bidirectional recurrent neural network with three hidden layers, adapted from [10]

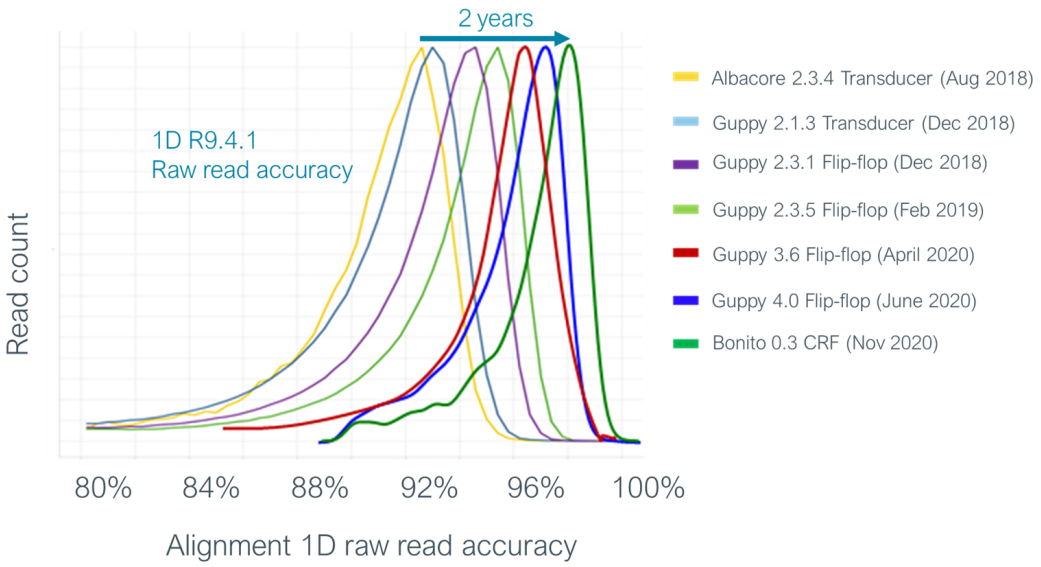

Figure 7: Comparing the accuracy of Oxford Nanopore base callers, adapted from [14]

collapse into one linear regression (Proof that mathematically!). Common activation functions include binary step function, sigmoid function, hyperbolic tangent or ReLU (Rectified Linear Unit) functions. For example, you could choose:

$$f(\vec{h}_{i-1}, \vec{x}_i) = \tanh(W \cdot \vec{x}_i + U \cdot \vec{h}_{i-1} + \vec{b}) \tag{1}$$

The weight matrices W and U and the bias vector $\vec{b}$ are parameters of the model. They need to be trained before the algorithm can be used. For that, a set of input vectors with known output has to be provided. It is very common to split the experimental data into two sets - a training and a test data set. Then the first set is used to optimize the parameters of your model and using the second set you can check the accuracy after training. Repeatedly adjusting the weights to minimize the difference between output and expected output is called backpropagation. Boža's base caller that was published in 2017 had an accuracy at around 77 - 89 % depending on the dataset and the training. Since then the accuracy of base callers that rely on machine learning methods increased considerably (Fig. 7).

For some really nice visualizations check out:
https://www.nanoporetech.com/how-it-works/basecalling

Experiment

The goal of this lab course is to find out which part of its genome the bacterium *Bacillus subtilis* has to replace in order to change its colony morphology. In the wet lab experiment, you will enable horizontal gene transfer (HGT) between *B. subtilis* and a different species, *Bacillus vallismortis*. In the data analysis part, you will analyse nanopore sequencing data to find out which genes have been affected by HGT and correlate them to colony morphology (Fig. 8).

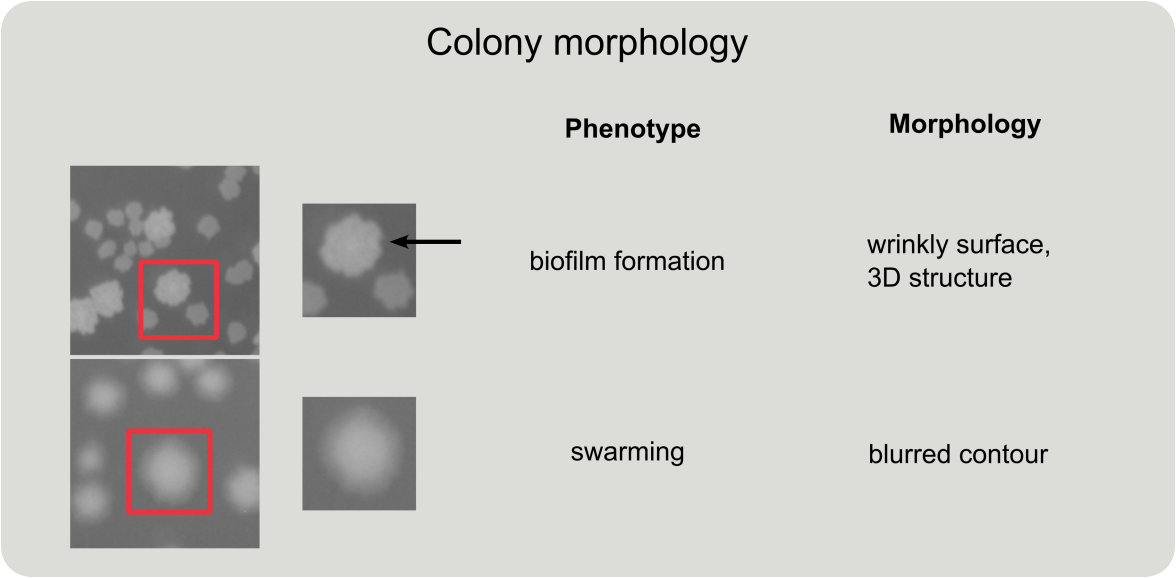

Figure 8: Some changes in phenotype of the cell result in a change of the colony morphology that can be recognized by eye. Cells that produce an extracellular matrix show a wrinkly 3D structured surface (biofilm formers). A blurred contour hints to cells with high motility (swarming phenotype).

The steps of the wet lab experiment are as follows. (i) Extract genomic DNA (gDNA) from the soil-dwelling bacterium *B. vallismortis*, (ii) let the closely related species *B. subtilis* take up the gDNA and integrate it into its genome via natural transformation, (iii) spread the hybrid bacteria onto an agar plate and (iv) identify and analyze hybrid clones that changed their colony phenotype.

In the morning, *B. subtilis* cells are diluted in fresh medium and incubated at 37 °C. Subsequently, genomic DNA of *B. vallismortis* will be extracted using the Quiagen® DNeasy Blood & Tissue Kit. The gDNA extracted with this kit primarily contains segments with a length of 30 kbp. In the meantime, *B. subtilis* escapes lag phase and enters exponential growth phase. (Please make yourself familiar with bacterial growth and the growth phases!) When growing exponentially, compentence for transformation will be induced by adding IPTG.

Step-by-step instructions:

- Centrifuge 1.5 ml *B. vallismortis* cells from overnight culture for 3 min at 16700 xg and discard supernatant or use cell pellet provided by supervisor.
- Resuspend cell pellet with enzymatic lysis buffer and incubate for about 30 min at 37 °C.
- Dilute *B. subtilis* overnight culture 1:100 in LB medium, then fill 4 wells of a 24 well plate (1 ml each) with the cells and 2 wells with only medium (blank).
- Use a plate reader to track the optical density at 600 nm to see when the cells escape lag phase.
- Liquefy LB agar and pour 5 LB agar plates.
- Let them dry open for 20 − 30 min, then close them.

———————————
—— Time for questions ——
———————————

- Finalize gDNA prep using the Quiagen® Blood & Tissue - kit.
- Measure the final concentration of the DNA you isolated using a photometer and calculate how much of it you need for your experiment.
- Add 600 μM IPTG (3 wells) and the amount of DNA you calculated (2 wells) to your cells. You now have: 2 blanks, 1 well where cells grow undisturbed, 1 well in which competence is induced but no DNA is added ("control") and 2 wells with competent cells and DNA. Let the cells take up DNA for 2 h.

———————————
—— Lunch break ——
———————————

- Fill 15 wells of a fresh 24 well plate with 900 μl PBS each.
- Dilute the 2 transformed + 1 control well by transferring 100 μl from well to well on the fresh plate.

- Plate a 10^{-4} and a 10^{-5} dilution from one of the transformed and from the control cells. From the second transformed well, only plate the 10^{-4} dilution. For plating, pipette 50 μl on an agar plate, add approx. 8 sterile glass beads and move the plate back and forth thoroughly. Then remove the beads and store the plate upside down in the plate incubator at 37 °C.

Next morning:

- Come to the lab to look at your plates and take pictures of them or let your supervisor sent you pictures. (Whatever you prefer.)

If you did the experiment on a Friday, do not incubate the cells at 37 °C but leave them on the shelves at room temperature instead. The cells will grow more slowly and proper pictures can be taken on Monday.

Analysis

Phenotypic Analysis

- Describe the different colony phenotypes you see on your plates. Give a rough estimate of the fractions. Do you also see them on the control plates?
- For your protocol, visualize the OD data measured during the transformation assay (excel sheet).

Genomic Analysis

Nanopore sequencing is performed in collaboration with the group of Dr. Paul Higgins at the Institute for Medical Microbiology, Immunology, and Hygiene of the university hospital Cologne.

- Familiarize yourself with the nanopore sequencing data you got:
  XX.fast5: raw data from Oxford Nanopore Technologies® MinION, they contain the current signal through the nanopore for every read. You do not have to work with this file unless you are interested in the raw data format
  XX_read_xx_signal.txt: gives you an example for such a current signal for one read extracted from the XX.fast5-file
  XX_read_xx.txt: is the translation of the current signal from XX_read_xx_signal.txt to bases produced by the supplied OxfordNanopore®basecaller.
  XX.fastq: contains all raw read signals from XX.fast5 translated to bases with information about the per-base quality and nucleotide sequence.
  XX.fasta: contains only the nucleotide sequences of the reads with no further information (simpler version of XX.fastq), in general: .fasta is a data format for nucleotide sequences
  dictionary\Bs166NCe.fasta: sequence of our *B. subtilis* strain, length: 4.21 Mbp
  dictionary\Bs166NCe.bed.txt: annotation of the Bs166NCe.fasta, giving you gene position, name and function for all known genes
  2blastn.fasta: contains four chosen reads you should align to the Bs166NCe.fasta using blastn
  blast2mut-folder: contains a python (.ipynb) and a matlab (.m) version of the same script that converts blastn results into a position/mutation list and a explanation file (HowTo.pdf) that helps you to use blastn and the python or matlab script.

- Use the **XX_read_xx_signal.txt**-file to plot the current signal of one read against time (in arbitrary units). Why is machine learning useful to analyze nanopore data?
- Extract the read length from your **XX.fasta**- or **XX.fastq**-file (Should be the same). Plot the lengths distribution. What is the longest read? What is the average read length? Does this compare to expected Nanopore read lengths? If not, why could that be the case?

Now, open the **blast2mut\HowTo.pdf** and follow the instructions.

- You get four files with lists of mutations and their positions. Write a short program to find mutations that are present in all four sequences (like this we get rid of sequencing errors).
- For those mutations (real mutations), plot the mutation density along the genome. Can you identify regions with high mutation density?
  Use the bed-file (**Bs166NCe.bed.txt**) to identify a gene/genes that are located in that region.
- Visit the SubtiWiki of the Stülke Lab in Göttingen (http://subtiwiki.uni-goettingen.de/) and look for the gene/genes you found to learn more about there functions/properties.
- Infer from the genomic changes which of the phenotypes you found on the plates was sequenced.

References

[1] D. Dubnau and M. Blokesch. “Mechanisms of DNA Uptake by Naturally Competent Bacteria”. eng. In: *Annual Review of Genetics* 53 (2019), pp. 217–237.

[2] B. Maier. “Competence and Transformation in Bacillus subtilis”. eng. In: *Current issues in molecular biology* 37 (2020), pp. 57–76.

[3] M. Förster et al. “Genome-wide transformation reveals extensive exchange across closely related Bacillus species”. eng. In: *Nucleic acids research* 51.22 (2023), pp. 12352–12366.

[4] E. L. van Dijk et al. “Ten years of next-generation sequencing technology”. eng. In: *Trends in genetics : TIG* 30.9 (2014), pp. 418–426.

[5] T. Hu et al. “Next-generation sequencing technologies: An overview”. eng. In: *Human immunology* 82.11 (2021), pp. 801–811.

[6] K. D. Pruitt, T. Tatusova, and D. R. Maglott. “NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins”. eng. In: *Nucleic acids research* 33.Database issue (2005), pp. D501–4.

[7] J. J. Kasianowicz et al. “Characterization of individual polynucleotide molecules using a membrane channel”. eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 93.24 (1996), pp. 13770–13773.

[8] E. A. Manrao et al. “Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase”. eng. In: *Nature Biotechnology* 30.4 (2012), pp. 349–353.

[9] I. M. Derrington et al. “Nanopore DNA sequencing with MspA”. eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.37 (2010), pp. 16060–16065.

[10] V. Boža, B. Brejová, and T. Vinař. “DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads”. eng. In: *PloS one* 12.6 (2017), e0178751.

[11] A. Graves and N. Jaitly. “Towards End-To-End Speech Recognition with Recurrent Neural Networks”. en. In: *International Conference on Machine Learning* (2014), pp. 1764–1772.

[12] H. Phan et al. “SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging”. eng. In: *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society* 27.3 (2019), pp. 400–410.

[13] P. Baheti. *Activation Functions in neural networks. 12 types & use cases*. Ed. by V7. 2021. URL: https://www.v7labs.com/blog/neural-networks-activation-functions#:~:text=drive%20V7's%20tools.-,What%20is%20a%20Neural%20Network%20Activation%20Function%3F,prediction%20using%20simpler%20mathematical%20operations. (visited on 10/26/2023).

[14] ONT. *How basecalling works*. URL: https://nanoporetech.com/how-it-works/basecalling (visited on 10/09/2023).

Paper to be prepared for the exam: [8]