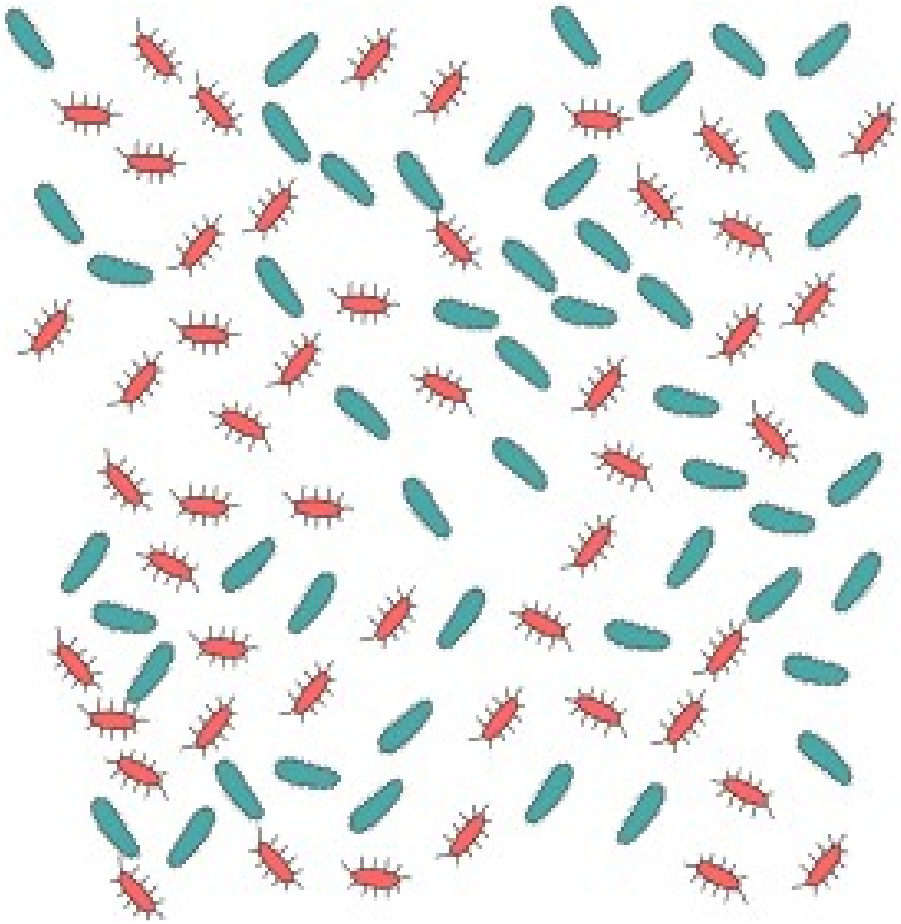
Maths and computer science in the study of microbes – bioinformatics, statistics, modelling

Mum! How can you be so certain that an apple a day keeps the doctor away?



Matti Ruuskanen, April Karkulahti, Leo Lahti

Department of Computing, University of Turku, Finland

Storyline

Because microbes are so small, humans have not even known of their existence until a few centuries ago. Studying microbes has been very different from how we study larger organisms because they cannot be observed easily in their natural habitats.

In the past, scientists mostly looked at microbes with microscopes and grew them in the laboratory. With these traditional methods we could see, for example, how microbes look and how they move, or if they produce specific chemicals or grow without oxygen. Much of our current knowledge of microbes comes from these experiments, but many details about how they live in nature were missing. Scientists have now started to study microbes in their natural habitats by analysing their genetic material, like DNA. DNA contains all the instructions for the biological machines needed by the organisms to live and grow. By studying these molecules, we can better understand what microbes do. By comparing DNA from different organisms, we can study their common ancestry and relationships. However, there is a lot of information contained even within a single organism, and there are massive amounts of microbes living in nearly every environment on Earth. For example, there can be billions of microbes from thousands of different species living in a single gram of soil.

Processing all of this information even from a single sample requires a lot of computer power. Luckily, computer technology has progressed greatly, and we now have the tools to process these mountains of information. Hence, many microbiologists need to learn and use math and programming. This science of studying information is called "data science", and it involves the use of computers and code, instead of microscopes and pipettes in a laboratory. By writing computer code, we can come up with new ways to understand the genetic code that the microbes need to thrive and reproduce in many different habitats, from arctic lakes to the human gut. Next, we show how data scientists think about microbes.

Microbes and math

1. ***Measuring the diversity of microbial communities.*** Currently, ecosystems around the world are threatened by human-influenced environmental change, like climate change, emissions of toxic chemicals and loss of *habitat*. These changes also affect microbial ecosystems. For example, global warming and increasing temperatures affect how many different microbes can survive in Arctic lakes.

Diversity is one of the most central concepts in biology. Diverse ecosystems contain many different species. Diversity of the microbial community is crucial for the normal functioning and health of many ecosystems and environments, like the human body. Therefore, microbiologists are very interested in measuring and comparing microbial diversity and have come up with many different ways to calculate it.

One way to measure diversity of a microbial community is to count how many different types of species live in it. The diversity of a community is high if the number of individuals belonging to different species is similar (=even), and the community is not dominated only by one or a few types of species. If the community consists only of a few species in uneven numbers, its diversity is low **(Figure 1)**. Scientists have also developed different diversity metrics which take both the number of species and their evenness into account, and these make the comparison of diversity easy.

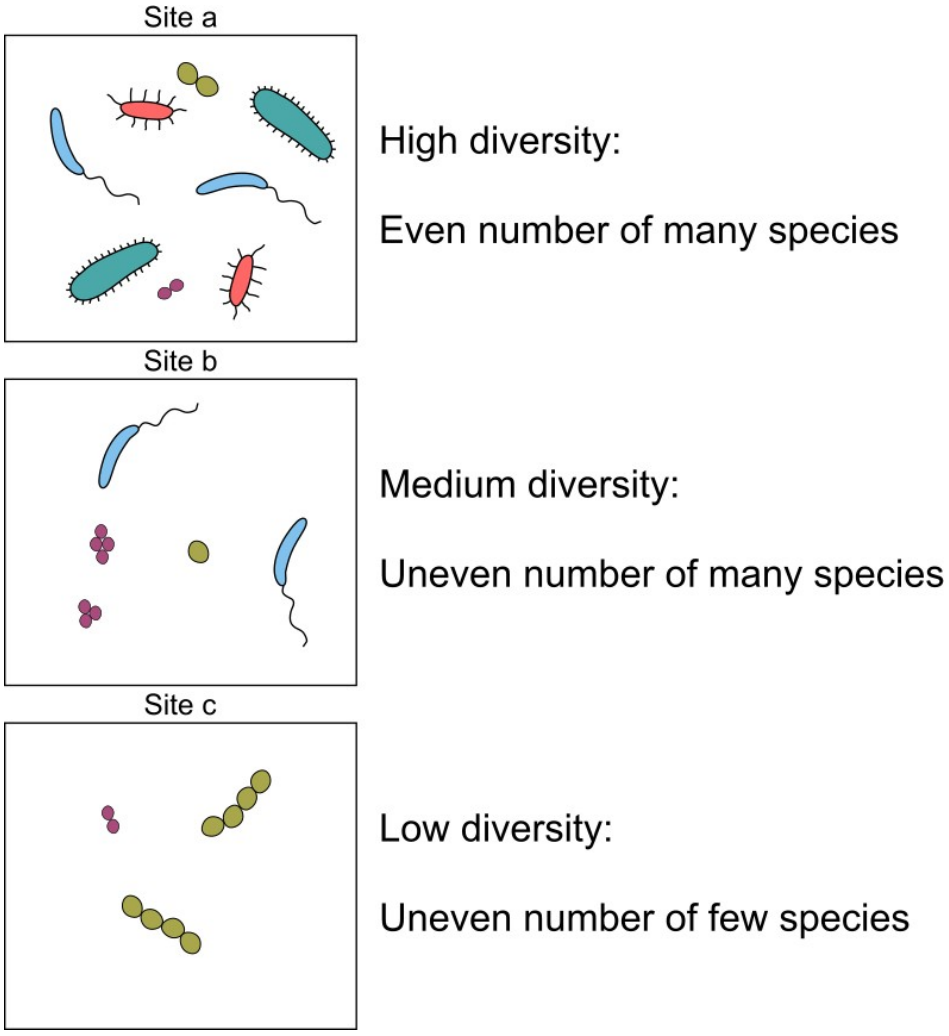


Figure 1. Three sites with communities showing different levels of microbial diversity. Note that despite their different diversities, each community still has the same number of bacterial cells. A simple way to calculate diversity is to count the number of different species that we can observe in a community. In this figure, the low diversity community has just two different species. Can you calculate how many different species there are in the medium and high diversity communities?

The diversity between two or more communities can also be compared by looking at how different these communities are compared to each other. The more of the same, or other closely related, microbial species these two sites have, the more similar they are. Through these kinds of comparisons, we now know that people or animals that live together tend to carry more similar microbes. Then we can say that there is a low diversity between you and your pet. The more there are differences, the higher the diversity.

2. ***Using data science to understand evolution.*** According to the theory of evolution, all known biological organisms on Earth share a common single ancestor, the first living cell. Evidence of this relatedness can be found, among other things, in the DNA sequences which microbiologists use to study microbes. The DNA sequences consist of the genetic code which can be written with 4 different letters. These letters are A for adenine, C for cytosine, G for guanine, and T for thymine. The chemical structure of a DNA molecule can be exactly defined by these letter sequences. When we analyse new DNA, we can use computers to find out how similar its sequence is to the sequences in previously-analysed DNA from other microbes.

a. *Central importance of ribosomal RNA gene sequences*

To study the evolutionary relationships between organisms, or their *phylogeny*, scientists can compare the sequences of their *genes*, or the DNA letters, that have gradually changed over time – mutated – as the organisms diverged. We can use mathematics to model this. Different genes change over time, or evolve, at different speeds. For a long time, microbiologists studying bacteria and archaea, the so-called prokaryotes, have been using a particular gene, the "16S ribosomal RNA gene" as a "molecular clock". This is an important gene that evolves at a rate which works quite well – not too fast, not too slow – for studies of the phylogeny of microbial species. (The equivalent gene in eukaryotes, the 18S ribosomal RNA, is used by biologists studying higher organisms.)

b. *Sequence alignments*. Once the DNA sequences of this gene have been gathered from all the different microbes in a sample, the letters of the DNA sequence can be aligned so that similar parts of the sequences are displayed side-by-side to simplify comparisons (**Figure 2**). Mathematical algorithms are used to construct these alignments and they can be visually examined to see which parts of the gene, or its sequence of letters, are different or similar between organisms.

c. *Phylogenetic trees*. Then another mathematical model is applied to the alignment to construct a phylogenetic tree, which shows the *evolutionary* distance between all the different species (**Figure 2**). This way we can measure the evolutionary distance between species, and it can tell how close (or distant) relatives two species are in the tree of life. Phylogenetic trees are quite similar to family trees, since both show the relationships between the subjects (family members or organisms, such as species) in the tree.

When we compare different microbial communities, it is important to not only look at how many different species there are, but also consider whether they are very close relatives, or if they diverged already a long time ago and are more distantly related. This is important because closely related species tend to perform quite similar functions in the microbial ecosystem, whereas more distant relatives can behave very differently.

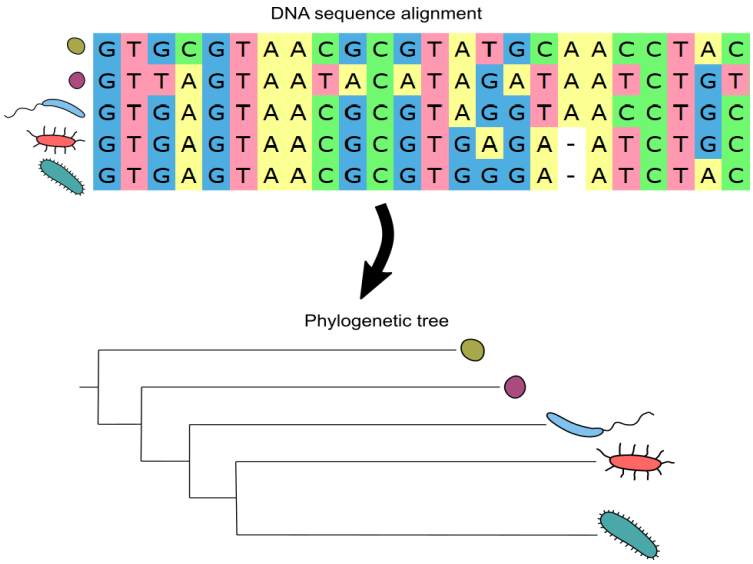


Figure 2. DNA sequences from different microbes (rows) are aligned so that similar parts of them are together (columns). The sequence of letters shows the structure of each DNA molecule. After the alignment, mathematical models are used to reconstruct the evolutionary relationships between the organisms which are visualized as a phylogenetic tree. Organisms which are on closely connected branches are more related to each other than those on distantly connected branches. The distance on the tree is a measure of their evolutionary distance, based on the alignment.

3. ***Relative abundance.*** High-throughput sequencing, the new technology used to produce massive amounts of DNA sequences, has revolutionized the study of microbial communities. However, it also has some downsides which need to be considered when analysing the sequencing data. One of these is that the size of the microbial community can vary across different habitats, for instance between my gut and your gut. The current analytical techniques do not tell us how many microbial cells are actually in a sample. Knowing the actual numbers of a microbe might be very important. For example, many pathogenic gut microbes are found normally in low numbers also in healthy individuals, but a high number of a single pathogenic microbe indicates an infection. Instead of accurately measuring actual numbers, with the sequencing methods we can only compare the relative amounts between microbes. These can be measured as fractions of the total (unknown) number of microbes, or percentages like 5% of *Enterobacteria*, and 95% of some other bacteria. This only tells if we see *more* of certain types of bacteria but it does not tell how *many* of them there are. When the data contains only this type of relative information, it is called *compositional data* (see **Box 1**).

Box 1. Compositionality

Comparing samples using compositional data is more difficult than with quantitative data, where the numbers can be simply counted, giving us *absolute counts*. For example, we might find that 50% sequences match the bacterium *Escherichia coli* in two samples we have analysed. If these numbers were absolute counts, we could say that both samples likely have an equal number of this bacterium. However, with compositional data we might end up with similar results from two samples which actually have very different communities.

In **Figure 3** we can see how the true (or absolute) number of the red species is ten times lower in sample 1 compared to sample 3. Unfortunately, with sequencing we can only measure the relative counts in these samples. The relative counts are exactly the same - 50% of the sequences are from the red species! This problem is caused by the simultaneous change in the number of the green species. Also, the relative counts of the red species are 10 times lower in sample 2 compared to sample 1. In this case, the true number of the red species stays the same, but the number of the green species increases 19 times! Because of these issues, many statistical techniques have been developed to understand such variation.

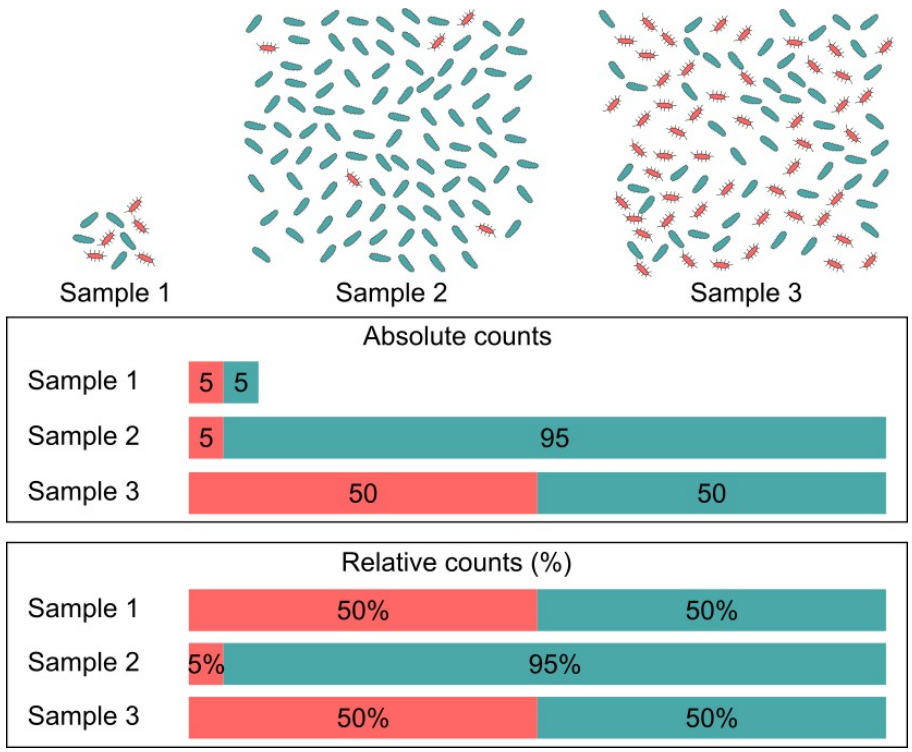


Figure 3. A comparison of the absolute and relative counts of two microbes in three samples. With the currently available sequencing technology we can often only measure relative counts. This can lead to false conclusions. Special statistical methods need to be used to make sense of this type of data.

4. ***How to compare: Statistics.*** Microbiologists are often interested in how microbial species, and the larger communities and ecosystems formed by these species, can be linked to different or changing habitat conditions. We can study how the temperature of the environment, or a disease like cancer, will influence the types and amounts of species that thrive in that habitat. Such changes in the environment may make life harder for the previous inhabitants, and bring in new members that are better suited to live under the new conditions.

As one example, scientists wanted to find out how the microbial communities vary between different parts of a hot spring. They first took samples near to the source of the hot spring where the water is very hot, and then further away from it where the water is cooler. They also recorded the temperature at each sampling site. Then they sequenced genetic material (DNA) from all samples, as described above, to find out which microbes were present in each sample. By combining the DNA and temperature data, they can analyse how these two components (parameters) may be linked, and if temperature influences the microbial community composition.

a. *What is the same? What is similar? What is different?* It is challenging to measure how microbial community varies with temperature because there are many other things that influence the community simultaneously, for instance pH, availability of sunlight, or nutrients. These and many other factors may also differ between different parts of a hot spring. In addition, there is a lot of natural variation even between two samples from the same location: two samples never have exactly the same species in the same amounts. If one water sample has 100 species, and the other has 101 species, are they really different? If not, how large must the change be before we can claim that there is a difference in microbial communities between hot and cold water? And how can we find out if temperature has any effect on this compared to

sunlight and all the other factors? Statistics provides important tools to study how the microbial community changes with temperature.

b. *Minimizing false and random results.* Because there are so many different microbes, it is difficult to directly compare each of them individually. By doing thousands and thousands of tests, some connections might also be indicated purely by chance! This issue is similar to throwing two 6-sided dice. If you throw them only once, you have a very low chance (about 3%) of getting both of them to land on a 6. However, if you throw them 10 times in a row, there is almost a ¼ chance to see this happening at least once. If you throw the dice 100 times, you are almost guaranteed to see two 6's at least once (about 94% chance). To combat this problem, microbiologists try to think carefully in advance what questions they like to ask, so that the possibility of false and random findings can be minimized.

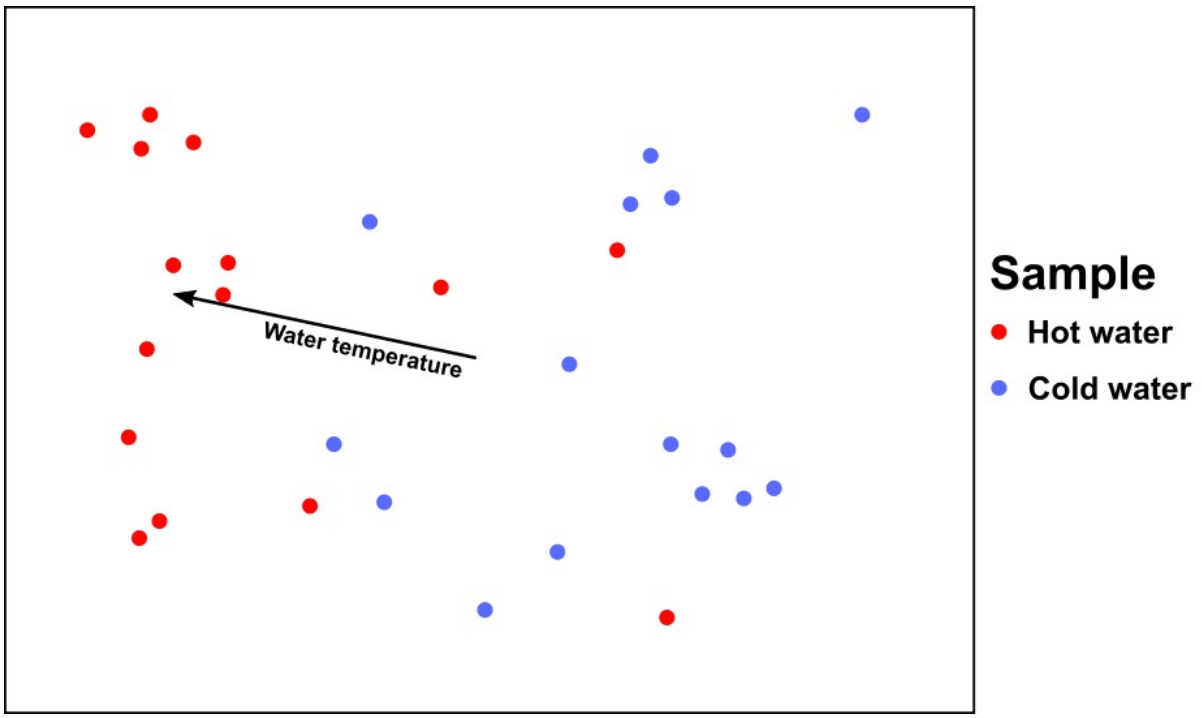


Figure 4. A map of similarities and differences between the microbial communities in samples from hot and cold waters near a hot spring. When two points are close-by on the map, this means that they have more similar microbial communities. The data on water temperature in the samples can be fitted on this map to analyse how strongly the microbial community changes by temperature. In this example, we see that the hot and cold water samples reside on different parts of the map, meaning that they carry different microbes. We also see that there is a lot of natural variation, and sometimes the differences are not so clear - perhaps because hot and colder water could be mixing?

Another way to simplify research is to look at similarity between the whole microbial community, instead of separately analysing every species. The similarity between microbial communities can be reduced to a map of distances (**Figure 4**); water samples with similar microbial communities are close to each other on the map in our example. With just a single test, a researcher can check if samples from different temperatures follow a specific direction on the map.

However, this way it might be difficult to find out which species prefer hotter water and which ones live in cooler conditions. But temperature is just one example. To test all interesting connections, the researchers might end up having to collect and analyse thousands of samples, if they can just afford to spend enough time around the hot springs!

5. ***How to compare: Modelling.*** In our previous example, the microbiologists might run into problems with traditional statistics. These statistical tests make simplifying assumptions about the connections. Because the microbiological communities are not simple, but very complex, it might be difficult to understand the real complex patterns with simple statistical tests.

Recently, new types of computer models that can help with this have been developed. These are called machine learning models. These models are mathematical algorithms which learn from their own mistakes. A good model can be trained by trial and error through many repetitions.

Such models can be used to make predictions based on new data. In the example given above, the model can also be used to see which microbes are most important for predicting whether the temperature of the sample is high (or low). The algorithms that are used in these kinds of models are very powerful and adaptive. They are so adaptive that, given the chance, they can memorize the data they used for learning and model it perfectly. For this reason, with machine learning modelling, the dataset is usually divided into two parts: a *training set* and a *test set*. The model learns when training on the training data and it is then tested on the test data. When testing a model, the scientists in our example feed the model only the microbial abundance data, and see if the model's temperature prediction matches the measured temperature. If the predicted and measured temperatures are close to each other, they know that the model works. Then it can be examined to see which microbial species are connected to high or low predicted temperatures. These types of models could be used to even diagnose diseases based on which microbes live in your gut.

Box 2. Health associations

Microbes live in all kinds of environments on earth. These environments range from oceans to soils, the air, and to the human body. When microbes interact through modifying digested food in the gut or chemically communicating with the immune system on the skin, they are also contributing to our health and disease. The new ways to study microbes in their natural habitat through DNA has also revolutionized our understanding of these connections between the human body and the microbes living in and on it. Using the various methods described above, scientists have discovered, for example, that the microbial communities in our gut are highly unique and are influenced by our genetics, where we live and what we eat. Many microbes, such as *Eubacterium rectale* and *Faecalibacterium prausnitzii* are known to be beneficial for the health of the colon. These bacteria ferment undigested material which has passed through the small intestine and produce butyrate, which is used by human cells in the colon as an energy source. In the past, it has been quite easy to identify the single pathogenic organisms which cause food poisoning and diarrhoea, like some strains of *Escherichia coli* and *Clostridium difficile*. However, more complex connections between commonly occurring microbes and chronic diseases, like fatty liver disease and inflammatory bowel disease have been discovered more recently. Also, many studies have revealed that the high diversity of the microbial community in the gut promotes health, and a low diversity seems to be connected with chronic diseases. These discoveries have often been made for the first time with the complex modelling tools and algorithms we have discussed. Aided by these tools, we are now discovering new connections, for example, between the gut microbes and the brain, and how the gut microbiome could be used to predict and diagnose diseases!

6. ***Universality of methods.*** In some cases, the methods used in microbiological research are highly specific to these problems. For example, it would be hard to find use for phylogenetic trees and assembly of genomes outside of the field of biology. However, algorithms used in the analysis of biological data are now very common also in other fields of study. It is possible that an algorithm used to model how a microbial community changes over time is also used to predict important events in geology, like earthquakes or volcanic eruptions, or crashes in financial markets. In these different cases, the algorithm remains the same, but the input data and the output of the models are different. The principles of data analysis remain the same in nearly all imaginable fields. Thus, in addition to the universality of math and algorithms, scientists working with these methods can apply their skills in a wide range of fields. Currently, it is not uncommon to find a physicist working in biology, or a bioinformatician applying their skills in finance.

7. ***The future.*** All these new tools and techniques have enabled us to study and understand the complex world of microbes better than ever before, but much still remains to be discovered! One of the most pressing limitations of most current sequencing methods is their inability to quantify actual microbial numbers. This issue is being actively addressed, and many of the new approaches also require advanced algorithms. Accurate quantification of the microbes would enable us to find new patterns and effects in complex communities. For example, when the picture of microbial communities connected to different human diseases becomes clearer, special algorithms could be developed to aid in their diagnosis. In the future, in addition to taking and analysing blood samples, microbial analyses of stool samples and swabs from skin could aid doctors in determining the health status of a patient. Recovery of patients from many diseases could also likely be improved by adjusting a treatment according to the composition of the patient's microbial communities. In addition, by fully understanding how the microbial communities in and on humans, animals and the environment function, we could be able to modify them to support their health. This could happen by introducing specific strains of new microbes and supporting the growth of already present beneficial microbes. The ability to precisely modify microbial communities could lead to further improvements in management of different environments, in agriculture, and in animal and human health.

Exercises

1. In what kind of environments can microbes exist? Make a list based on your previous knowledge and what you have read.
2. How can microbes affect your life? Write down as many ways as you can think of.
3. What challenges have researchers faced when studying microbes? What are the current challenges in the study of microbes?
4. Arrange the following steps of scientific study to correct order.
    a. Collecting the samples
    b. Publishing the results
    c. Coming up with a study question and hypothesis
    d. Sequencing the samples
    e. Statistical analysis of the data
    f. Interpreting the results
5. Read the sections 'How to compare: Statistics'"and 'How to compare: Modelling'. Design your own study question and write down a study plan (how you could perform the

research). Also mention if you expect specific results from the study (what is your hypothesis?). Some examples:

a. Design a trial to test the adage 'An apple a day keeps the doctor away'.

b. How does the microbial community change over time on a rotting fruit?

c. How does the microbial community change with the distance from the toilet bowl in a public toilet?

d. How does the microbial community differ from the top part of the soil compared to the deeper parts of the soil?

Glossary

Microbe: A microscopic organism, which is so small that it cannot be seen with the naked eye. There are many different types of microbes: bacteria, archaea, eukaryotes, or viruses. Most organisms with diameters under 0.2mm (or 200µm) are usually thought to be microbes.

DNA: Abbreviation for DeoxyriboNucleic Acid. This is the hereditary material in all cellular organisms. DNA molecules are found inside living cells and contain the genetic code of the organism, or its "building instructions" that can be passed to offspring. These molecules are also durable outside of living cells, so DNA from dead cells can be found in many environments.

Ecosystem: The collection of all living organisms and their surrounding environment in a defined place. Examples of very large ecosystems would be the Amazon rainforest, or the Pacific Ocean. Examples of smaller ecosystems are a park and a lake. Also, very small ecosystems exist, like the microbial ecosystem in a yogurt can in your fridge.

Habitat: The natural living environment of an organism. The habitat of an organism is a place where it typically lives and grows. The organism survives in and can get all or at least most of its basic needs fulfilled inside of its habitat. For microbes this can mean, for example, suitable temperature, moisture, acidity, and oxygen levels, and availability of light and nutrients.

(Bio)diversity: All living organisms in nature and their variety and variability. The (bio)diversity of an environment or an ecosystem can be studied and measured by counting all different organisms living in it. Highly diverse ecosystems have many different types of species and they are often resilient to disturbances, so biodiversity should be protected and maintained.

Evolution: The gradual changing of biological species over time. According to the theory of evolution, natural selection acts on the individuals of a species. This means that not all individuals of a species survive or are otherwise able to reproduce and pass their own genes onward. Only the genes of those individuals which are best adapted to their environment, or habitat, are passed onward to their offspring. This leads to a constant adaptation and change in all biological species.

Phylogeny: The relationships between biological organisms. All cellular organisms on Earth are thought to originate from a single living organism, which lived about 4 billion years ago. This microbe is called "Luca", for the Last Universal Common Ancestor. The amazing diversity of life has developed since then through the processes of evolution and speciation, or formation of new species. Thus, all currently living organisms are related, or have common ancestors, and their relationships can be studied and described.

Gene: A piece of genetic material, like DNA, which carries the information on how to construct a specific biological molecule, such as a protein. In humans, differences in a single gene can, for example, influence your eye colour or the shape of your earlobes.

Algorithm: A series of instructions. In computer science, algorithms are written with a coding language, and they instruct the computer to do a specific task. Algorithms can also be described

in normal written language. For example, "take an apple from the tree and put it in the basket" is an algorithm. This would be a useful algorithm for picking apples. However, you might run into problems if there are no apples left in the tree, or some of them are rotten, or your basket is too small. Thus, algorithms can be very simple, but often they have to be quite complex.

Microbiome: The collection of all microbes (bacteria, archaea, eukaryotes, viruses), their genetic material, and their surrounding environmental conditions. In macroscopic (animals, plants, etc.) contexts, a similar term "biome" is used to describe the collection of both biological and non-biological factors in an environment.