# IT & Biology

DNA (short for deoxyribonucleic acid) carries the genetic information. It consists of a pair of nucleotide strands that wind around a common axis to form a double helix. Nucleotides are the basic building blocks of DNA and consist of a sugar residue (deoxyribose), a phosphate group and a base (one of A, C, G and T). For computers, nucleotide sequences are simply character sequences (strings) over an alphabet of four letters.
Here is a short example:

GATTACATTACAGA
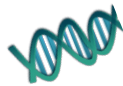
When scientists discover a new organism, they usually try to find the closest relatives among the known organisms to learn something about the new organism from them. For example, when a new disease-causing kind of bacteria is discovered, it is very likely that medicine that works against the closest relatives of the new bacteria also work against itself. In contrast, developing new medicine takes time, and if there are people ill already, they need help fast.
To compare two organisms based on their DNA, the nucleotide sequences are adjusted so that similar or identical nucleotides are arranged vertically to produce matches. Here is an example of such an alignment of the sequence above and CTACCAGTACACAC:

Hi! My name is Anna-Lena Lamprecht. I am an assistant professor in the Software Technology group at the Department of Information and Computing Sciences. I teach courses about programming and software development, and my research is about developing new programming and software construction methods that make it easier for scientists to create the software that they need in their research themselves. Therefore I work together with a lot of people from other disciplines, for example geoscientists and life science researchers. For me the various possibilities for collaborating with and learning from other disciplines is what makes the information and computing sciences so interesting.

Curious what life science research would have to do with information and computing sciences? Try to solve the example problem from bioinformatics described below.

GA**T**TA**CA**T**TACA**G**A-
-C**T**AC**CA**G**TACA**C**A**C

To judge how "good" an alignment is, a basic score can be computed by simply counting the matches (the positions where the nucleotide is the same in both sequences) in an alignment. With this the alignment above has a score of 8. The higher the score, the closer is the relationship between the organisms.

? **Suppose the sequence above is from a new, dangerous bacteria X. Can you identify the closest relative based on their sequences as shown below?**
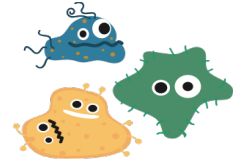
Bacteria A:
GGAGTACAATACCG

Bacteria B:
CGATTAGATAACAG

Bacteria C:
TTACAGAGATTACA
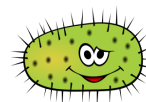
Answer!

Bacteria B is the closest relative. The best alignment has a score of 11.

-**GATTA**C**AT**T**ACAG**A
C**GATTA**G**ATA****ACAG**-

Note that the DNA sequences of real organisms are much longer. For example, the human genome consists of more than six billion nucleotides! And also the algorithms that are used to compute alignments and scores from large sets of real sequences are more sophisticated. Good that software programs can to all this automatically...

Also interesting: On **tolweb.org/Life_on_Earth/1** you can explore the "tree of life" interactively. To find the relationships between the organisms shown there, sequence alignments also played a role.