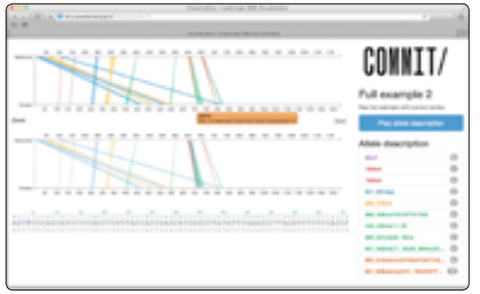
9. Rapidly finding variations between human genomes



We have developed a software tool, called the Variant Description Extractor, that rapidly compares one human genome with another in order to find small but crucial genetic differences. Our tool generates a complete description for the human genome in about four hours.
The human genome contains twenty to twenty-five thousand genes distributed over a long molecule, called DNA. Genes can be described by long strings of the four letters A, C, G and T. Each of them stands for a simpler molecule in the DNA. On average, humans only differ 0,1% genetically from each other. However, especially for finding causes and solutions to diseases it is crucial to find and understand these small differences.

ICT science question

The main scientific challenge is twofold. First, how to calculate short and unique descriptions from long strings of the letters that compose the genes? The genes can hold thousand to many millions of these four letters.
Second, how can this calculation be done within an acceptable and minimal amount of computational time?

Application

Our Variation Description Extractor is integrated in the Mutalyzer suite. Mutalyzer is a very popular web-based software tool primarily designed to check descriptions of sequence variants according to the standard human sequence variant nomenclature of the Human Genome Sequence Variation Society. Mutalyzer aims to encourage the proper use of nomenclature in publications and reduce redundancy in gene variant databases. This greatly improves the findability of variants. Ultimately, these genetic variant descriptions are used in the diagnostics of hereditary diseases.

Jonathan K. Vis
jvis@liacs.nl
www.mutalyzer.nl

COMMIT/ project

Alternative Application

Although our software tool is now focused on comparing DNA-strings composed of four letters, it can easily be used on strings composed of other signs. The tool might for example be applied to describe the differences between two natural language texts. The fundaments of the tool and the algorithm do not have to be changed. Certain operators that are now especially designed for DNA-comparison can easily be disabled.

Nice to know

The Variant Description Extractor has already been downloaded 66 million times.

The Mutalyzer suite enables medical researchers to find all variants and combinations of variants in DNA sequences quickly. These variants can be used to effectively research complex diseases.

We use a state-of-the-art algorithm to quickly find variants in DNA sequences. The algorithm is supported by a Python suite of compatible tools adhering to the well-known HGVS standard.

The efficient extraction of variants on a individual level enables the targeting of diseases on a person-to-person basis. Making it a prerequisite for and a first step towards personalized medicine.

We propose an efficient algorithm for the extraction of biological meaningful descriptions of the variance between complete genome sequences with minimal length and computation time.

This research was supported by the Dutch national program COMMIT/