Teaching an algorithm to identify cancer from sequence data

Deep learning has revolutionised cancer research. Deep neural networks can automatically detect features within a patient's sample data that can be used to identify cancers. In the future, learning algorithms will be able to identify potential early-stage cancers from a blood sample. Esa Pitkänen and his research group at the Institute for Molecular Medicine Finland (FIMM) are developing a new generation of deep-learning algorithms.

Algorithms have been used to identify cells in sectional images of tissue samples. For instance, if tissue cells appear atypical, the algorithm will spot this and determine if the cells are cancerous. DNA sequence data from tumours is now being used along with imaging data to identify cancers.

"Until recently, it was difficult to tell from a DNA sequence what kind of tumour an identified sequence came from. Now new technologies and deep learning algorithms have been created," Pitkänen says.

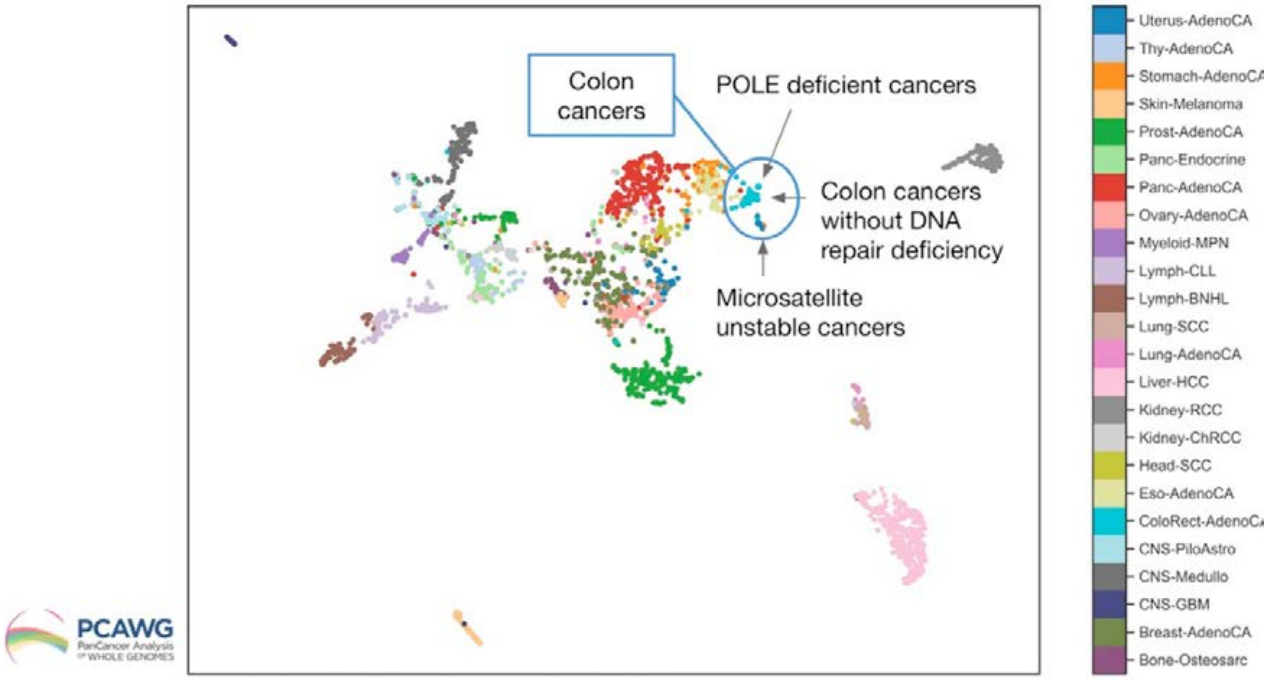
Pitkänen and his team are developing algorithms that identify short, repetitive snippets of DNA sequences. These algorithms can be used to find DNA sequences that mutate frequently in a particular type of cancer or to which certain proteins involved in gene regulation bind. Analysis of these sequences can be used for various purposes, such as charting the causes of cancer and developing medicines.

"The replication of DNA in cell division is not perfect; mutations can occur during the process. The division of a single cell involves the copying of about six billion nucleotide pairs in DNA, so errors will inevitably occur. Even the slightest probability of errors is enough to guarantee mutations," says Pitkänen.

"If enough mutations occur in genes that prevent tumour growth, for example, cancer may start to develop."

An example of this is a point mutation in which one base within the DNA strand is replaced with another. The enzymes involved in copying DNA may make a mistake when a cell is dividing, for instance by incorrectly repairing the part of DNA that was damaged by ultraviolet radiation from sunlight. A typical mutation caused by ultraviolet radiation that can result in skin cancer is that two consecutive cytosine (C) base molecules in the base pairs of human DNA are converted to two thymine (T) base molecules. When skin cancer-specific mutations of this type are detected in sufficient numbers, the algorithms can learn to associate them with a particular type of cancer.

"We try to predict the type of cancer and tumour from the mutations. This also provides information on how treatment can be developed."



See also Jiao *et al. Nature Communications* 2020

In their work, Esa Pitkänen and his research team used one of the world's largest data sets, from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project. It contains 47 million mutations. The data is from sequenced tumour samples from 2,600 patients. The collection included 37 tumour types from different cancers, including colorectal cancer, lung cancer and melanomas. Prima Sanjaya used deep neural networks to create a machine learning model that processes the sequencing data of each patient and presents this data in two-dimensional map format. In this image, each point represents one distinct tumour obtained from a patient. The different colours represent the different types of tumours. Interestingly, the model groups colorectal cancers together, but also distinguishes three subtypes (marked with arrows in the figure).

Identifying cancer from blood sample DNA using algorithms

Pitkänen and his group analyse blocks of sequences and train algorithms to pinpoint deviations in them. From these abnormalities, the algorithm can detect tumours and classify them into different cancer types.

"Before joining the Institute for Molecular Medicine Finland, I worked at the European Molecular Biology Laboratory in Heidelberg, where I participated in the Pan-Cancer Analysis of Whole Genomes (PCAWG) project. This project involved the analysis of more than 2,600 entire cancer genomes. My group is using data from the PCAWG project in several of our cancer genomics projects."

An algorithm developed by the group has been taught the mutations found in the tumour samples of 2,600 cancer patients. This sample set contained about 47 million mutations. Approximately 50 million somatic mutations were found in the sequence data.

"We trained the algorithm to try to deduce the type of cancer from these sequence changes. Once the algorithm is given all the mutations of different tumours and their sequences, in the future it will be able to determine the kind of tumour that has been detected. This deduction process is based on the algorithm learning these connections."

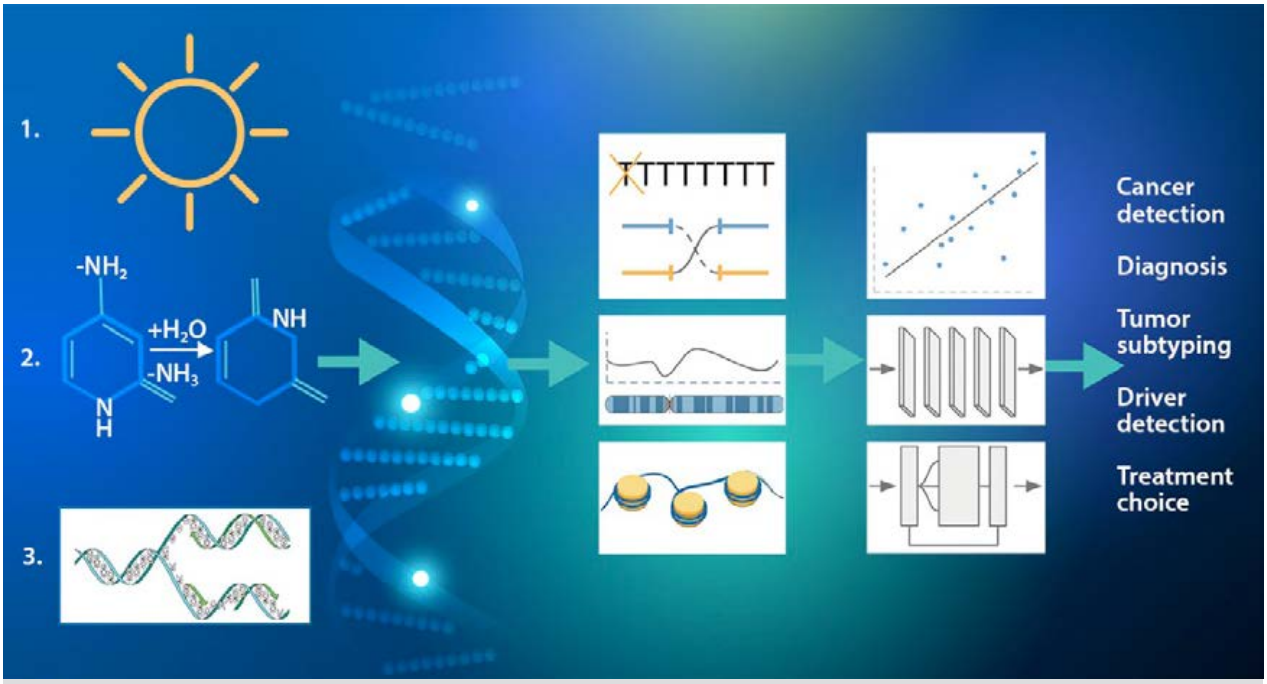
Through deviations in the sequence data in tumours, the algorithm learns to identify when a given tumour corresponds to a particular type of cancer. It can group tumours based on sequence data alone.

"A researcher in my group, Prima Sanjaya, has developed neural network models for analysing sequence data. Every now and then, researchers come across cases in which a cancer has metastasised without being able to tell where it has spread from. Such cases could in the future be dealt with by means of a liquid biopsy – that is, it will hopefully be possible to determine from a blood sample if the patient has cancer, and if so, what kind."

Liquid biopsies are based on the fact that cells release into the bloodstream and other body fluids a type of DNA called cell-free DNA (cfDNA). Cancer cells also release DNA, which makes it possible to test for cancer mutations in the blood plasma.

"If a liquid biopsy shows traces of cancer, we don't know exactly what kind of cancer it is, as it could have entered the bloodstream from anywhere in the body. If we have the means to examine these cases more closely, for example with deep learning algorithms, we could obtain valuable information on where in the patient's body the tests should be focused. The algorithm may indicate that the source of the cancer DNA could be the large intestine, for example. I believe that such algorithms will become extremely important. Liquid biopsy and algorithms can make it possible to diagnose cancers without surgery."

Sources of mutations



1.External factors (e.g. UV radiation from sunlight), **2.Internal factors** (e.g. a spontaneous deamination reaction, in which the amine group of a base changes, for example from adenine to uracil) **3. DNA copying errors.**

A mutation is a change in the nucleotide sequence of DNA or RNA. A nucleotide consists of a base, a sugar molecule and a phosphate group. The sugar in DNA is deoxyribose and the sugar in RNA is ribose. The four nitrogenous bases that DNA contains are guanine (G), adenine (A), cytosine (C) and thymine (T). RNA has three of the same bases as DNA, but instead of thymine its fourth base is uracil (U).

A mutation may be a change of a single nucleotide – that is, a point mutation – or the change may involve multiple nucleotides. In a point mutation, one base is replaced by another in the RNA or DNA strand. Large mutations can involve thousands of nucleotides, and are called structural changes. A structural change can affect multiple genes at the same time. Cancers are usually caused by several somatic mutations. Mutations of this kind are not inherited, and can occur at any time of life from embryonic development onwards. Mutations may bring about a change in the functioning of a normal cell, causing it to begin to divide uncontrollably.

At the middle of the picture are presented different types of mutations, distribution of mutations on chromosomes and epigenetic information. Epigenetic inheritance is influenced by many external factors, such as nutrition. An example is the development of identical twins so as to become distinct from each other in appearance.

Modelling mutations:

Linear models

Deep neural networks

Transformer models. Transformers are a family of deep learning models that work particularly well with certain types of data, such as textual data. This makes them well suited to machine translation, for instance. In cancer research, transformer models can draw attention to mutation types that are important for identifying a particular type of cancer. For example, in skin cancers that contain many sunlight-induced mutations (C> T, CC> TT), the transformer will focus on these particular mutations.

Identifying intestinal cancers using algorithms

In addition to hereditary factors, the development of cancer also depends on the person's lifestyle. Plentiful research has been conducted at the University of Helsinki on various cancers, such as intestinal cancers.

"It is known that eating red meat contributes to the incidence of colorectal cancer. The mechanisms by which the disease is caused require further research, but in recent years a lot of progress has been made in understanding the significance of DNA alkylation reactions, which are caused by red meat, for example."

Colorectal cancer is one of the most dangerous cancers in Western countries. In countries such as Finland, it leads to death in 30 per cent of cases. About 15 per cent of colorectal cancers belong to the cancer group that exhibits microsatellite instability (MSI). Microsatellites are sequences of DNA that can vary in length from person to person, and thus function as individual identifiers in much the same way as fingerprints. Microsatellite instability occurs when the post-replication repair mechanism of cellular DNA does not function properly. This causes mutations to begin to accumulate, especially in microsatellites.

"In an MSI tumour, microsatellites are easily vulnerable to single-base additions or deletions. For example, out of eight consecutive adenine microsatellites, one adenine may be lost. When it occurs in a gene, such a change causes a complete transformation in the content of the amino acid chain of the protein that is encoded by the gene. If there are enough changes in genes that are important for preventing uncontrolled cell growth, cancer may begin to develop."

MSI is often associated with other cancers as well as colorectal cancer, such as stomach, uterine, ovarian and brain cancer. MSI analysis can be used in the prognosis of cancer. The treatment choice may be influenced by the analysis.

"An interesting thing is that the deep neural network is also learning to classify different subtypes of cancers. For instance, it identified the MSI subtype of colorectal cancers," says Pitkänen.

The ELIXIR Node in Finland, hosted by CSC – IT Center for Science, is one of the main partners in the Personalised Medicine in Europe (PerMedCoE) project. For example, the three-year the HPC/Exascale Centre of Excellence for Personalised Medicine in Europe project is aimed at making effective use of cancer-related data in healthcare and speeding up the process of diagnosis

"Individualised treatments of the future, among them cancer treatments, will be based on a precise understanding of the patient and their illness. This will result from gathering a large volume of data of different types, such as tumour-related data and imaging data during cancer treatment. Many data collection methods produce a mass of data, and the new computational methods developed for analysing it require a very large amount of computational resources," says Pitkänen.

"Developing a new computational method from the idea stage into a functional healthcare technology is a huge challenge in an operating environment like this. With cancer treatments in particular, it is important that information relevant to patient care be made available to the doctors as rapidly as possible. I'm confident that the results of the PerMedCoE project will provide a basis for deriving relevant information from a colossal volume of health data to help doctors in their work, and thus significantly improve treatment outcomes."

Ari Turunen

MORE INFORMATION:

HPC/Exascale Centre of Excellence in Personalised Medicine (PerMedCoE)
https://permedcoe.eu

Institute for Molecular Medicine Finland (FIMM)
https://www.fimm.fi/en/

CSC – IT Center for Science
is a non-profit, state-owned company administered by the Ministry of Education and Culture. CSC maintains and develops the state-owned, centralised IT infrastructure.
http://www.csc.fi
https://research.csc.fi/cloud-computing

ELIXIR
builds infrastructure in support of the biological sector. It brings together the leading organisations of 21 European countries and the EMBL European Molecular Biology Laboratory to form a common infrastructure for biological information. CSC – IT Center for Science is the Finnish centre within this infrastructure.
http://www.elixir-finland.org
http://www.elixir-europe.org