Fighting Cancer with Mathematics

Extensive data sets and databases are increasingly being used in cancer research. The research group of Sampsa Hautaniemi, Professor of Systems Biology at the Faculty of Medicine of the University of Helsinki, develops methods that can be used to integrate data from various sources, such as DNA, gene expression and protein function. When the analysis results are combined with biomedical databases, it becomes possible to generate experimentally testable predictions. This is useful in the diagnostics and design of treatment methods, for example.

Sampsa Hautaniemi worked at the Massachusetts Institute of Technology (MIT) before setting up his own research group at the University of Helsinki in 2006. Hautaniemi's laboratory analyses complex, disease-related biological systems using mathematical methods. The analysis of data masses is not possible without computational assistance.

"Biomedical research requires databases and computational methods, especially in the interpretation phase of the results", says Hautaniemi.

The objective of the systems biology group, which operates at Biomedicum, is to apply computational methods to medical research questions. For example, which genetic profiles affect cancer risk or what is the prognosis of a patient with a particular genetic profile? The aim is to find a unique treatment in accordance with the genomic profile for the patient.

"Our goal is to understand the behaviour of the cancer cell and look for targets that, when their activity is modified, allow cancer cells to be destroyed with minimal side effects. When wanting to treat a cancer patient, you must first understand how the tumour cells make decisions on how they grow, multiply and move. We pursue this through genome-wide measurement and mathematical methods."

In the treatment of breast cancer, for example, it is important to be able to predict the probability of metastases emerging. Even though the treatment prognosis for breast cancer is improving all the time, metastases greatly increase the risk of disease.

"The problem is that we do not know how and which cells detach from the tumour, where they go and how they function there."

The aim is to deduce who has a high probability of forming metastases by studying gene activity and combining data. Current measurement methods, such as microchips and new generation sequencers, generate enormous amounts of data.

"At this time, we do not yet know the main internal cell factors that affect the treatment response of cancer. That is why we use methods from different levels that measure the whole genome in research."

In addition to DNA and RNA sequencing, such methods include, for example, epigenetics, or analysing the impact of lifestyle on gene function. Proteomics, which determines the function of proteins and their structure, is also important.

Suitable medication based on data

More than four billion observation points can be measured from one cancer tumour. From this mass of observations, you should be able to identify the most characteristic factors for cancer development and drug response.

According to Hautaniemi, there has been quite a change compared to the situation 10–20 years ago when the usual number of observations to be processed was a few dozen or hundred.

"In addition, databases have genome-wide data available on thousands of cancer patients. Utilising this data alongside Finnish material is important, but challenging."

In addition to prognosis, Hautaniemi's group also looks for suitable treatment methods based on computational analysis. Hautaniemi's group is mapping, for example, the impact of genetic modifications on drug response. Cytostatics, which destroy cancer cells, are used in the treatment of cancer. It is important to find a suitable cytostatic because the patient does not always respond well to the given drug.

In cooperation with the group of Professor Olli Carpén, Hautaniemi's laboratory has used genome-wide data on hundreds of ovarian cancer patients in their research. The researchers have been looking for subgroups of patients that have developed a resistance to conventional chemotherapy in which platinum derivatives and taxoids are used as cytostatics.

The research project uses hundreds of thousands of processor hours of supercomputer computing time and dozens of terabytes of storage capacity.

"For a person with a certain type of genetic profile, some medications may even be harmful, while others provide the optimal benefit."

How data becomes knowledge

Hautaniemi and his group have developed methods by using data related to lymphoma together with the group of Professor Sirpa Leppä. The challenge is how to convert the data collected from genes and proteins into knowledge. Observations from clinical samples are always rather noisy and multidimensional, meaning that there are thousands of genes, proteins and potentially interesting areas of DNA. Therefore, it is essential to answer the correct and necessary medical questions so that the results are useful. The research questions can then be solved by mathematical methods.

When analysing lymphoma and ovarian cancer data, Hautaniemi's group used the so-called deep sequencing method. The method involves DNA or RNA being divided and sequenced, after which the base sequence of the molecules is converted into a format understood by a computer. There may be hundreds of millions of short sequences converted into a computer format. According to Hautaniemi, when converting medical data into knowledge, the most significant bottleneck that is faced is the comprehension of medical questions so that they can be modified into computational problems.

To solve this problem, Hautaniemi and his group have developed a software program called GROK (Genomic Region Operation Kit). It allows questions to be converted into computational problems and solved based on the data. GROK is a universal tool and it has been used to understand the progression of prostate cancer. The study was conducted in cooperation with the laboratory of Professor Olli Jänne. The cooperation resulted in a better understanding of the function of the FoxA1 protein with the AR protein, which is the main protein affecting prostate cancer. Furthermore, the study found that a large number of FoxA1 proteins provide a poor prognosis and a small number provides a good prognosis. In future, the results can be used to prepare a treatment prognosis and for planning treatment. According to Hautaniemi, the methods developed can be applied to any kind of cancer.

"We have used the methods we have developed to study, for example, breast,

prostate and ovarian cancers. Although the tumours are found in different organs, they have a significant number of similarities at the molecular level. Therefore, in future, it might be possible to use a breast cancer drug for certain subtypes of ovarian cancer, for example. Prior to this, it must be possible to characterise the subtypes of each cancer. This means that, in future, we will be able to reliably find similar cancers regardless of their location and then recommend effective medication suitable for them."

Hautaniemi believes that cancer cell sequencing will be part of routine cancer diagnostics in future.

"We are striving to find the factors for each tumour type and individual tumour, and it is only a matter of time before we understand the biology of tumours so well that we can quickly calculate a prognosis and combinations of drugs that are likely to be effective based on their genome. Computational sciences play a key role in achieving this and utilising technology."

ELIXIR: European assistance for the processing of biomedical data

These days, the amount of data produced by life science experiments doubles every few months, and the amount is still growing. The experiments also produce a completely new kind of data. The accumulation of huge amounts of data from research has created a need to systematically manage all that information. The objective of ELIXIR is to harmonise data storage, processing and analysis.

In many respects, databases are starting to be vital for life science research, but they have often been maintained alongside other research activities and dependent on fixed-term research funding. One of the main objectives of ELIXIR is to secure the funding of the most important databases containing biological research data. When the system compiling and distributing information is permanent, research groups can build their own operations on it. The ELIXIR infrastructure also provides a system and funding pathway for services developed in Finland that are significant for the whole of Europe. Everyone does not have to produce the same database on their own; instead, data that has been created once can be used effectively in multiple locations, and tasks can be shared.

"The field of bioinformatics is so vast that no single laboratory can provide all services. What the Finnish and ESFRI project infrastructures bring with them are a certain clarity and an improved flow of information. We know what is being done and planned elsewhere", Hautaniemi says.

Ari Turunen

FURTHER INFORMATION:

Genomic region operation kit
http://csbi.ltdk.helsinki.fi/grok/

Ovaska, Lyly, Sahu, Jänne, Hautaniemi (2013):
Genomic region operation kit for flexible processing of deep sequencing data

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