Bioinformatics to revolutionise healthcare

Efficient data processing speeds up diagnoses and enables personalised drug treatments.

Data obtained from human genes and proteins enables quick diagnosis of illnesses and more detailed and personalised patient treatment. Those at risk can be screened better and more effective medication can be chosen once the patient's genome is known. The challenge lies in how data is processed and where it is stored.

Owing to new data analysis methods and improved computing power, the processing of gene and protein data can be accelerated from several days to less than half an hour. But in order to do this, data must be pre-processed, that is, redundancies removed, and be quickly and securely available.

Bioinformatics measurement methods have developed apace, producing huge amounts of data. Now we are in a position to understand the operation of an entire biological system. This means that genes – or the proteins created by them – are studied simultaneously. These 'omics', if you like, include determining DNA sequences (genomics), computer modelling of protein structures (proteomics), study of genes in cell tissue (transcriptomics) and metabolic products (metabolomics). Such methods can be used to study molecular interaction and find signs of changes in the human body to identify diseases.

"We can already do a great deal, but combining genetic and clinical data requires a lot more data storage and computing power, and all this in a secure way," says **Katja Kivinen**, Head of the FIMM Technology Center from the Institute for Molecular Medicine Finland (FIMM).

Screening for disease risk

The proper pre-processing and modelling of data are requirements for future research, with a view to providing more accurate disease prediction and even personalised 'silver bullets' in terms of medication. Kivinen gives two examples: pre-screening of disease risk factors, and drug treatments.

Nature Medicine ran an article in spring 2020 based on data analyses made in the Finngen project. On the basis of genomic data, Professor **Samuli Ripatti**'s research team at FIMM was able to identify a Finnish population group that had a 60% probability of suffering from cardiovascular diseases or diabetes at some point in their lives. The research material consisted of 135,000 Finnish people providing a blood sample. Data obtained about individual risk factors was

combined into a genome-wide risk score.

"Once a person's genetic background is known, clinicians will be able to offer preventive measures and treatments to improve health as well as save time and money," says Kivinen.

"In future, population-wide screening can be customised. For example, some women are now screened too early for breast cancer given their small inherited risk, while others are screened far too late. If we use genomic data to take account of inherited risk factors, women could be invited for their first mammography at the optimal time, thereby maximising the chances of discovering cancer at an early stage while minimising unnecessary radiation from repeated screening."

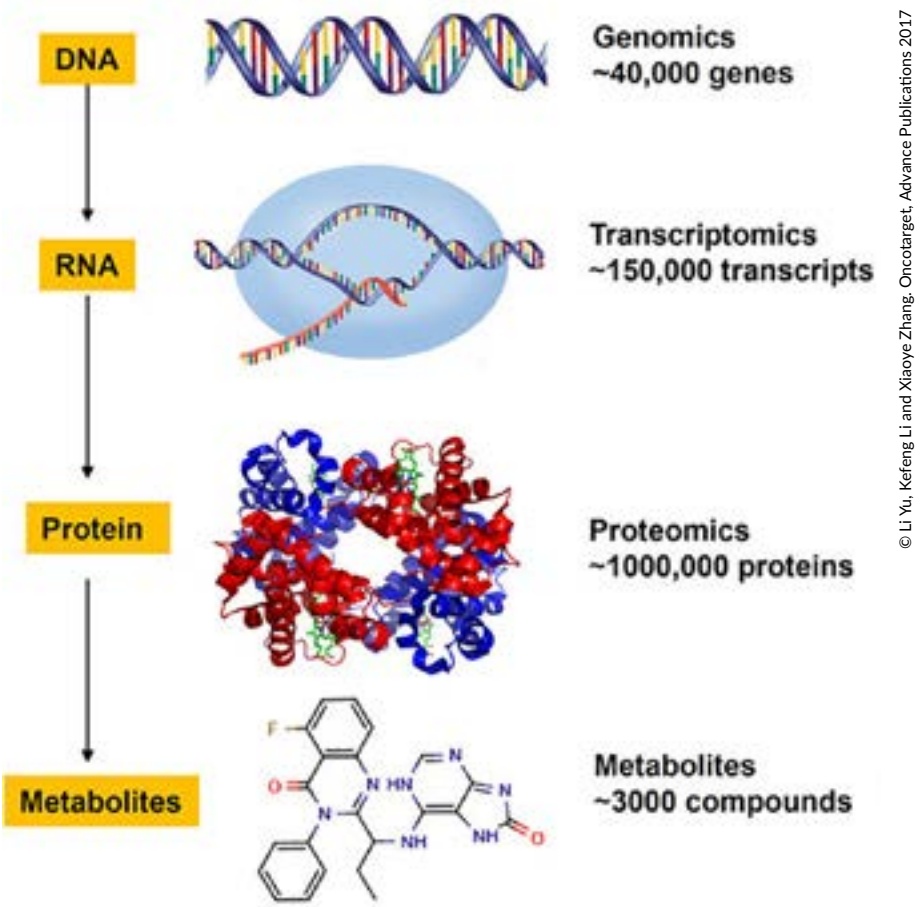
Personalised drug treatments arriving

People react to drugs differently: some do not get the necessary dose, while others suffer from side effects. The reason for varying responses may be our physical characteristics, other medication, or genome. Through the pilot phase of 'HUS e-care for me' project, we are optimising treatment for patients with leukaemia and other blood cancers. The project kicked off in summer 2019 and uses artificial intelligence to combine patient-specific biological data produced at FIMM with information about cancer type and status to identify suitable, personalised treatments and stop cancer progression.

"Sometimes a drug may not be effective, or have lost its effect. Cell cultures are made from leukaemia patients' blood samples, in order to analyse which combination of drugs works best."

Genome and transcriptome sequencing are performed on the same blood sample. A transcriptome provides information on whether the operation of genes has changed as a result of genetic mutations.

"If a drug no longer has the desired effect, we can find out what kinds of mutations have occurred. Are some genes working, or not, as a result of a mutation? And how do mutations affect metabolic pathways? Now we can see which drug is the best suited for different blood cancer patients directly from blood samples."



In the '+1 million genomes' project, the genome of one million Europeans will be sequenced by 2022. In human genome, there are about 20,000—25,000 known protein—coding DNA genes and in the proteome there are about 250,000—1,000,000 proteins, while in the human body there are thousands small molecule metabolites.

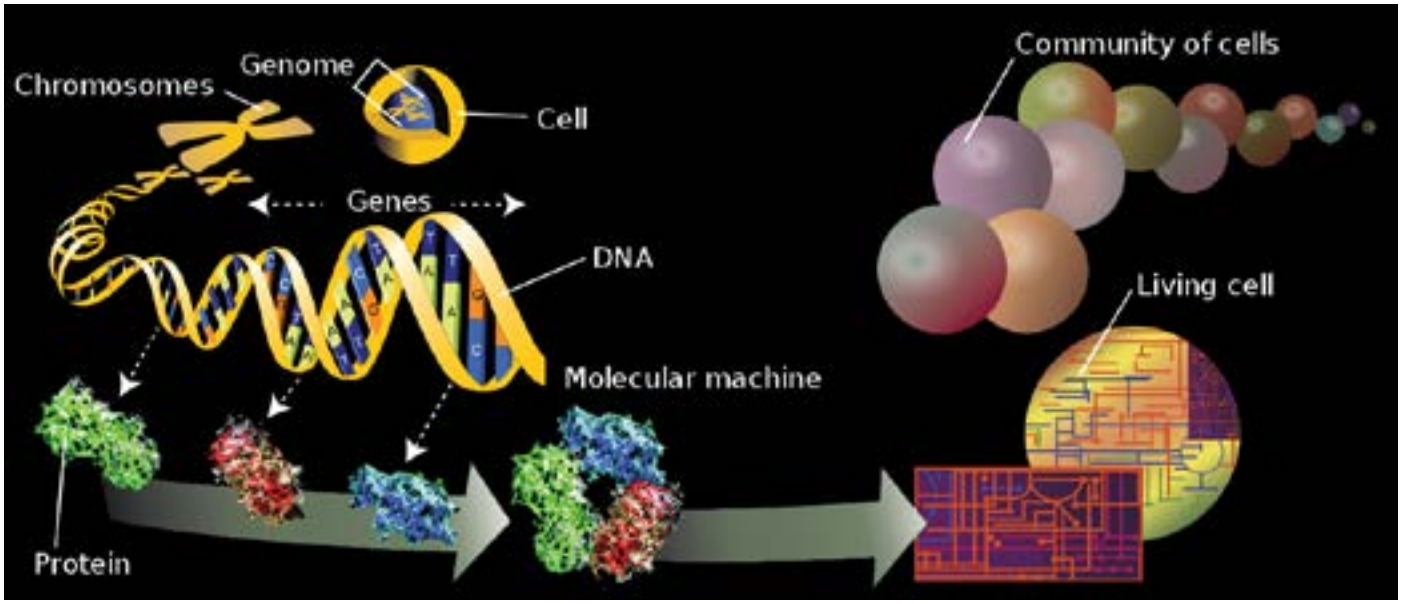
Disease diagnosis much quicker

Personalised drug treatments are possible if enough data is available on the patient, and it has been stored and pre-processed correctly. Alongside eleven other academic and commercial players, the Finnish ELIXIR Center CSC and the Barcelona Supercomputing Center (BSC) kicked off the European HPC Center of Excellence for Personalised Medicine (PerMedCoE) project in October 2020. The project is developing algorithms that can significantly reduce the computing time required for analysis. Analysis of genetic and protein data is becoming faster, facilitating and speeding up disease diagnosis and identification of the appropriate treatments. A genome analysis can take weeks or even months. Thanks to super computation and better software, diseases can be diagnosed in future within just hours or days.

Projects like this are important for research teams at FIMM who are routinely working with massive amounts of data.

"The amount of data is increasing at an accelerating pace owing to more powerful equipment and methods," says Katja Kivinen. "At the moment, we are chronically in short supply of data storage space, and data pre-processing takes too long to enable us to clear up job backlogs and send analysed data to research teams. A secure environment for storing and processing data is vital when dealing with human data. Commercial cloud services offer a secure operating environment, but are too expensive for most researchers. What is more, some data requires a carefully tailored pre-processing and analysis environment and is poorly suited to the options available in cloud services."

Data processing can be alleviated by dividing work between CSC and FIMM. In the



Genomics according to the description by William Crochot.

pilot stage, genomic data is transferred from FIMM to CSC on a superfast and secure optic cable. Data pre-processing and quality assurance of analysis are fast, because the data is located at CSC. In future, CSC will distribute data back to research teams on a national basis.

"Previously, it took us 2–3 days per human genome to determine what kinds of changes had taken place. Thanks to this cooperation, the manufacturer of our sequencing equipment has given us access to an optimised computing server that can compress the processing of one genome into 20 minutes. This will help us deal with the backlog of genomic data processing at FIMM and free our bioinformatics specialist to perform other work – such as planning and facilitating various data integration procedures."

CSC has developed a new interface to its Allas -service for genomic data from FIMM, to be used by Finnish research teams. The research teams will receive a message once their genomic data is ready, and transfer it to their own project area in CSC's ePouta environment. Following the pilot stage, the portal's operating principle will be offered on a larger scale to all research teams in Finnish universities producing 'omics' data.

"The interface is vital for us, as data volumes are increasing and data security requirements are becoming tighter, and it is increasingly difficult for us to maintain a data storage and processing environment at FIMM. We must start moving more and more raw data and process data to CSC and make it accessible for research teams."

Another important area of development, according to Kivinen, relates to data storage and related imaging services.

"Image processing is usually done on the server connected to the actual instrument, which contains the necessary processing software. Owing to slower transfer speeds, moving processing to a cloud service is not always a viable alternative, at least not in all parts of the country. So image processing may also occur on location in the future but, like genomic data, processed data should be shared through CSC."

Ari Turunen

MORE INFORMATION:

https://www.fimm.fi/en/

https://www.cleverhealth.fi/fi/ecare-for-me

https://ec.europa.eu/digital-single-market/en/european-1-million-genomes-initiative

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

ELIXIR FINLAND
Tel. +358 9 457 2821s • e-mail: servicedesk@csc.fi
www.elixir-europe.org/about-us/who-we-are/nodes/finland

ELIXIR HEAD OFFICE
EMBL-European Bioinformatics Institute
www.elixir-europe.org