Research-appropriate genetic data on the Finnish population exists fragmented all over the world in various databases and data storages with varying arrangements. Therefore, there is a need to create a domestic, secure service for the management of Finnish data that would cross organisational boundaries, be network-based and well-coordinated. By coordinating the data in different locations in just one place, the data could, with the permission of the owner, be released for legitimate purposes, such as research, product development and medication.

The human biology is very complicated, more complicated than previously thought. The expression, structure and function of genes and the building blocks of the body, or proteins, require advanced mathematical, computer science and statistical methods, i.e. bioinformatics.

New ways to study and prevent diseases are constantly being discovered through bioinformatics methods, such as gene sequencing. DNA sequencing is the starting point where the order of the four bases – adenine, guanine, cytosine and thymine (A, G, C, T) – within a DNA molecule is determined when deciphering the genetic digital code. Each ACGT base is a similar nugget of information to a computer bit, zero or one, which, as a long chain, contains the instructions for a programme.

Sequencing methods have improved and become cheaper, and this has significantly increased the possibilities of biology and medicine to produce this kind of data. The data is now being used to find out what digital messages have been written on the molecules of life for the survival of organisms.

However, data is only the first step towards interpretation. The interpretation of digital genomic data, that is, how the information stored in the genome manifests itself in the body, is still under development. In the last ten years, for example, researchers in Sweden have been creating a map (HPA Human Protein Atlas) on how genes are expressed as proteins in different cells and then combining this information with microscope images of cells. This allows you to see which gene is expressed in any given cell and is involved in the development of proteins and, hence, larger structures, such as neural fibres, hair follicles or light-sensing structures in the fundus of the eye. However, a clear, deeper level map on

Storing the Whole Genome of the Finnish Population? The Data Will Benefit Disease Research

Extensive research projects are being conducted on Finnish genetic heritage and genomic data is being produced and analysed all the time. However, the national objective is to store the data produced on Finnish people in Finland, allowing analysts to combine the data with other health information. The utilisation of genomic data in healthcare is still in its early stages. Data analysis offers many opportunities for companies in the bio-industry, including in Finland.
how molecules that are operating on a nanometre scale produce these functional, microscopic structures does not exist yet. The structure of each cell requires millions of molecules in cooperation. The building instructions stored in genomes and the resulting molecules form a self-organising network that current research tries to understand.

Finland is fairly well positioned to be an international actor in the management of genomic data, but there are too few experts in individual organisations. The data masses required to understand genomic data are large and the analysis requires specialised expertise that does not sufficiently exist in Finland yet. There is a need for cooperation in genomic data management and for more interpreters specialising in data. Finland will gain more expertise once the creation of a framework for storing Finnish genomes is achieved. Initially, this would mean a national reference database created from the data of tens of thousands of people. It would be beneficial for diagnostics, such as in improving medical treatments, as it is already possible to determine, for example, suitable and safe medication based on the patient’s genomic data.

Good organisation of data facilitates disease research

Analysing data from molecules, cells or whole organisms requires that the data is well organised. The data produced with sequencing, microscopes, mass spectrometry or computer simulations must have common file standards and sufficient machine-readable interfaces to be followed when the data is stored. A good indicator of the degree of data organisation is if another research group can utilise the data as well as its original producers.

When data is well organised and described, it can be combined. Combining supplementary information, such as a prescription, genome and long-term treatment results, is a prerequisite for developing a deeper understanding.

Data organised in the hands of skilled analysts will help achieve breakthroughs in research. The US company GRAIL, for example, seeks to understand the underlying mechanisms of cancer. The earlier the cancer is detected significantly improves the prognosis of the disease.

The GRAIL project has involved the collection of samples from 10,000 patients and their consent for the analysis of the diverse data created from the samples. The idea is to use the cancer tumours of this group of patients to create a database against which blood samples can be screened.

Cancer tumours are usually the result of a change in the genome of a cell of the person carrying the disease, making the cell abnormal. At the cellular level, each cancer is a rather unique disease that looks like its carrier; what they have in common is the reckless growth of abnormal cells. Cancer utilises the normal regeneration and healing mechanisms of the body to selfishly spread its own genetic instructions. The genomes and the digital information contained therein of two humans are, on average, 99.5% identical. That is why the progression process of many cancers is well-known despite the individual nature of cancers. Consequently, it is justified to study how changes in individual or multiple nucleotides (ACGT) in the genome affect the balance of the cell’s molecular network so that the cell becomes a cancer cell.

In the GRAIL project, millions of unique changes in genomic data that may cause cancer are sequenced from the genomes and cancer tumours of patients. The project will create a database that allows health care professionals to detect early stages of cancer, even directly from the bloodstream. If the innovation is successful, cancer screening can be started earlier, meaning that the tumours are still microscopically small and easier to manage through, for example, medication.

Conducting similar research in Finland is possible by combining health and genomic data. The Finnish ELIXIR node, for example, has already started building the secure infrastructure necessary for the management and storage of genomic data.

Understanding the emergence of diseases at the molecular level

Services provided by ELIXIR Finland

There are hundreds of times more data on the information contained in DNA available for science than ten years ago. Understanding of how the information stored in the genome is transmitted at the molecular level, for example, to proteins, and further to three-dimensional functional units of cells, is growing at a rapid pace. When human biology is understood from the cellular level to the molecular level, it improves quality of life and the treatment of diseases.

One of the most important research subjects in bioinformatics is understanding the underlying mechanisms of diseases. The functional unit encoded by a gene is a protein. It is a chain of hundreds of units, or amino acids. There are 20 different amino acids. The protein chain guided by genes becomes a functional unit of the cell, such as an enzyme, only after it has folded into its three-dimensional state and can start interacting with other molecules in the cell. An incorrectly folded protein can lead to illness because it does not function as expected in the network formed by molecules important to life.
Sometimes, for example, there is a change in the genetic code at a critical point for the folding of this critical functional unit, or protein. Cells self-modify the composition of the proteins created, and thereby their structure and function. This may correct the error in the genetic code. On the other hand, what may also happen is that the protein breaks down in the cell’s own process. Most diseases can be traced back to situations where a biochemical reading error has occurred in an important part of the dynamics of the cell’s molecular network. On the other hand, this may just be a variation that only results in dietary recommendations to the person in question. The effect of molecular level changes on the data stored in the genome depends on many things, as DNA includes a “backup” of each gene from both parents. There are even several versions of some genes. Even though the logic and knowledge on the main players in the network of biological processes are pretty much accounted for, the dynamic entity cannot yet be understood, let alone predicted or modified medically, as well as desired. Predicting the risks of contracting coronary artery disease, for example, has become more accurate thanks to the data obtained from the genome, but the understanding of molecular level events is at a stage where the components are known but there is a struggle to understand their interaction or defects occurring at the molecular level. However, molecular level understanding of diseases means more accurate and earlier diagnoses, that preventative measures can be initiated early and that those at risk, for example, may choose to change their lifestyle.

Tommi Nyrönen
Ari Turunen

FURTHER INFORMATION:
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