Mapping the Genomes of all Organisms Enables the Development of New Vaccines and Medicines

With the development of the methods used in bioinformatics, also the costs have lowered. It has become faster and cheaper to find out the genome of various organisms. However, we have a formidable task to be done to understand the information contained in the genome of various organisms and humans. It will require cooperation between various research organisations and well organised databases.

The mapping of the whole human genome was completed in 2003. Owing to the Internet, the Human Genome Project was completed earlier than anticipated, since it enabled efficient cooperation between various laboratories. The entire human DNA was sequenced. The human genes have been packed into three billion base pairs. Now, the next step is to find out how these genes work. Through the analysis of the base pairs of the genome we will begin to understand the pathogenetic mechanism of various illnesses and effective forms of treatment.

Today, research is generating quite versatile genome data. The aim is, for example, to use the information to evaluate the status of the environment and effects on health by analysing microbes, to cultivate edible plants into plants that will better withstand draught to alleviate the crises caused by climate change, or to develop drugs against diseases for which there is no cure at the moment. To do this, new kind of linking and analysis of the sources of data will be needed.

All the genomes of the known species will be mapped

It is becoming faster and cheaper to find out the genome of various organisms.

Probably one third of all terrestrial species are found in the Amazon area.

Now, as part of the Earth Bio-Genome Project (EBP), the aim is to map the genome of all eukaryotic organisms. Eukaryotic archaea and eubacteria, i.e. prokaryotes, are cells the DNA of which is constituted of only one chromosome. The group of eukaryotes consists of unicellular protozoans and three groups of multicellular organisms: plants, fungi and animals.

By means of bioinformatics, we can map the remaining 80 to 90 per cent of those organisms whose genome still remains unknown. In 2011, Census of Marine Life estimated the number of animal species to be approximately 8.7 million, 6.5 million of which are terrestrial and 2.2 million are marine animals. According to the estimates based on high-performance sequencing methods, there may be as many as 5.1 million species of fungi. There are approximately 400,000 plant species.

For the first time in human history, we will have the opportunity to efficiently sequence the genome of all known eukaryotic organisms. EPB's aim is to sequence all of the known 1.5 million eukaryotes. Samples are being gathered all around the world. Part of them, probably around half a million, will be derived from botanical gardens. The rest will need to be directly collected from the nature. One of the most significant collection sites is the Amazon. In January 2018, EPB launched cooperation with a Brazilian gene bank project which concentrates on the organisms of the Amazon area.

The Amazon area has a richer variety of plant and animal species than anywhere

Researches developed an antihypertensive drug from the venom of jacaraca viper.

Ocean Sampling Day.

else in the world. Probably one third of all species are found there. Rain forests are the home of a huge potential of new drugs.

For example, ACE inhibitor, i.e. the angiotensin-converting enzyme, was discovered from the venom of the jacaraca viper in the Amazon. The enzyme generates angiotensin, which helps lower blood pressure and lighten the pumping of the heart. In the 1970s, researches developed a synthetic version of the venom of this snake.

Massive Data Archives

The oceans are the largest continuous ecosystem in the world. The significance of planktons for the global climate is at least as important as that of the rain forests. However, only a fraction of those organisms which create this ecosystem, have been classified and analysed. The ecosystems constituted by planktons contain a vast amount of life: in every litre of ocean water there are more than 10 million organisms, containing viruses, prokaryotes, unicellular eukaryotes and cnidarians. These genuine organisms contain bioactive compounds which can be used in the pharmaceutical industry, foods, cosmetics, bioenergy and nanotechnology. In 2009-2013, the researchers of Tara Oceans, an international expedition, collected 35,000 biological samples in 210 different measurement locations from oceans around the world. This is the largest plankton collection until this day. Another campaign in which samples were collected from the sea, was Ocean Sampling Day. In that campaign, research stations were asked to collect samples and to generate data. BioSamples collects descriptions and metadata from biological samples that have been used in research. The samples are references or have been used in various databases.

Analysing genomes and the proteins that determine their operation is a huge task, which would not be possible without cooperation. The European life science infrastructure for biological information ELIXIR provides an efficient platform for cooperation with members from nearly 200 research organisations, and an infrastructure which is used by almost half a million researchers. ELIXIR enables access to various data archives.

Massive sequencing of cultivated plants and forest vegetation allows us to do research on what is causing plant diseases. EURISCO (European Search Catalogue for Plant Genetic Resources) contains information on 1.9 million cultivated plants and their wild cousins. The samples have been collected by nearly 400 different organisations. A total of 43 countries are involved, and the aim is to preserve the agrobiological diversity of the world.

Uniprot (Universal Protein Resource) is collecting protein sequences and annotation data. An annotation means the determination of the functioning of the protein on the basis of the sequence. Owing to Uniprot's data, we can learn more about the functioning of proteins and their interaction with other molecules as well as their location in cells and organisms. The aim is to collect all publicly available protein sequence data. Uniprot is the largest publicly available protein sequence database.

The European Nucleotide Archive ENA is a collection which offers free access to all published nucleotide sequences and annotated DNA and RNA sequences. The International Nucleotide Sequence Database is a collaboration forum between DNA Data Bank of Japan (Japan), GenBank (USA) and ENA. New data is synchronised between these three databases every day. Already in 2012, these databases contained the entire genomes of 5,682 organisms. The amount of data is doubled every ten months.

The European Genome Archive EGA is one of the largest public data storages in the world with patient data from biomedical projects. EGA stores the genotype and phenotype data collected from humans on the basis of a separate consent for research use of the sample and the data. Thanks to EGA, many of the ELIXIR research projects have become possible.

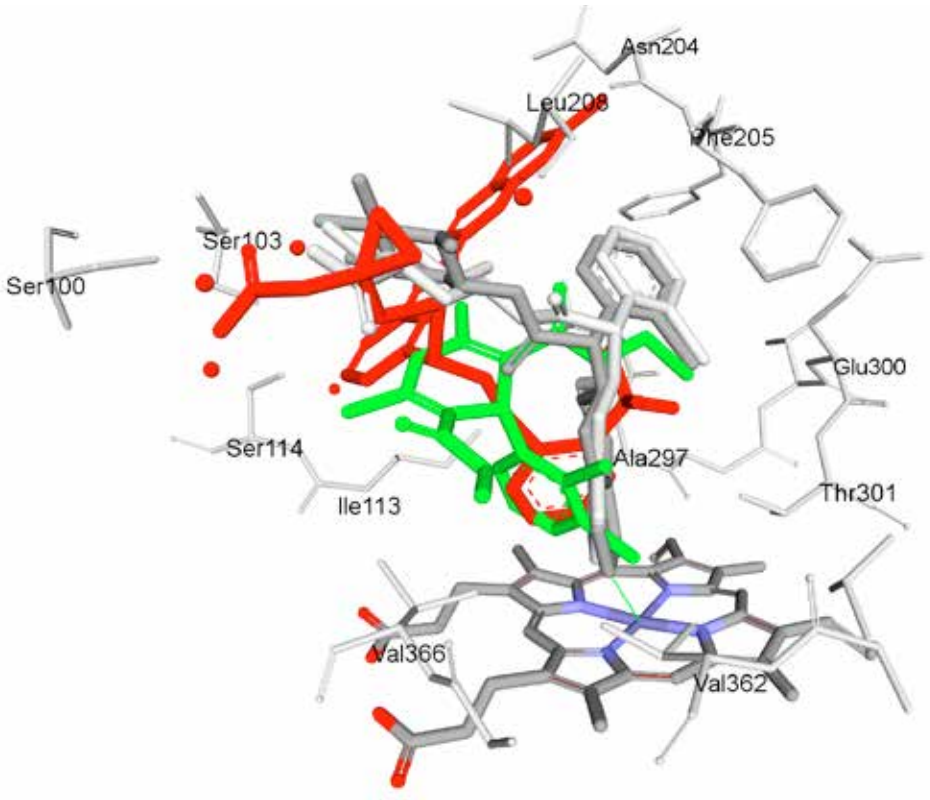
Biomedical Data to the Health Records

The ELIXIR infrastructure has more than 20 member states. Biomedical data is offered for use by researchers through the national centres in the member states. The benefits are indisputable. The genes of dogs and cats have proven useful in the analysis of rare human diseases. Through the Finnish centre the researchers have had access to a DNA bank of dogs and cats, the data of which has allowed us to discover the gene of a nerve degeneration disease, for example. The aim is now

to develop a drug for the disease. Canine genes have proven useful in the research of human diseases, because the canine and human genomes are 95 per cent identical. The canine gene bank contains more than 70,000 samples from 60,000 dogs and 300 breeds of dogs. It is probably the largest of its kind in the world.

According to the estimates, by 2025 we will be able to sequence 100 million to two billion human genomes. To receive the best benefit from the data, genotype data should be linked to other health data. ELIXIR will be able to do this. The research infrastructure consists of nearly 200 organisations which form a federation, a network of trusted parties, which enables secure processing of human data. By 2016, with the help of the ELIXIR infrastructure, 21,000 scientific articles had been drawn up and 8,500 patents had been granted. The patents had been applied for vaccines, biomarkers, enzymes and prevention of the Ebola virus.

The order of magnitude of a single atom of a living biological molecules is one tenth



"To receive the best benefit from the data, genotype data should be linked to other health data."

of a nanometre. Should one carbon atom of that biomolecule be of human size, it would mean that its functioning would have a crucial impact on events that take place dozens of millions of kilometres away.

The diameter of our solar system is of the same order of magnitude.

If only one carbon in the biological molecule is replaced with another atom, say nitrogen, it could serve as a decisive property for the drug, whether it works or not, for example. Just this particular atom could be the one with which the drug molecule is making an attempt to attach to a protein, but fails to get a strong enough grip on it as a result of this change.

The protein, which the drug was supposed to influence, again, forwards orders to other proteins in our cells. If influencing the order is left undone, influencing on the biological message chain is left undone, too.

We could also ask whether all parts of the message chain located in the cell are flawless. All these factors will have an impact on whether researchers will be able to design a drug molecule correctly so that it can help the cells heal.

Unlike in space, there is no vacuum in a cell. The cells are full of constantly interacting biomolecules.

Our chances to have an impact on the fusion reaction of the sun, for example, are much more limited than the impact of the atom level digital information stored in the living molecules on people falling ill, even though the difference in the order of magnitude is the same.

Tommi Nyrönen
Ari Turunen

MORE 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

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

www.elixir-finland.org

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