cOmpRession of Genomic dAta to facilitate precision MedIcine

Results in Brief

New technique helps to tame the deluge of genomic data

Technology now makes it feasible to sequence the whole of the human genome, bringing the prospect of personalised medicine a step closer. A French tech start-up has been testing its algorithm for extreme data compression which could make storing and transferring the data easier.

HEALTH

© Sergei Drozd, Shutterstock

High throughput sequencing – a technique which can run millions of sequencing processes in parallel – makes it possible to sequence the whole of the human genome in around 1 day.

This technology has not only cut the cost of sequencing but also brought the possibility of personalised medicine much closer. Being able to see someone's unique molecular and genetic profile could help scientists predict how likely they are to develop a given disease. It could also help doctors choose the best treatment when an illness develops.

But genome sequencing produces an enormous amount of data. "If you sequence the whole genome for just one patient, you are talking about 3 billion base pairs. If you want to sequence a molecule to find some variants, you have to do it several times so it might end up being over 30 times that amount," says Jennifer del Giudice, CEO of Enancio and project coordinator of EU-funded project ORIGAMI.

How to store the data

“So the big question is – how do we store and transfer this critical, personal data? It might need storing for 10 or 15 years. It’s a great opportunity to treat disease in a different way, but how do you handle the information?”

Enancio has developed an algorithm called Lena, based on an idea of the company’s CTO Guillaume Rizk and designed to serve the genomic data sector, which offers strong compression with no loss of data.

Lena proves a high compression ratio, compresses and extracts data quickly and with no loss of data and requires fewer computing resources to run than other solutions on the market.

Through ORIGAMI, the Enancio team tested the performance of Lena for these metrics on the latest version of the Illumina platform – the most commonly used provider for high throughput sequencing worldwide.

The results showed Lena can reduce the size of data by a factor of five in respect of the generic compression software currently used. This offers substantial savings in terms of data transfer times and the cost of storing data.

Five times smaller

“People already applying compression techniques will bring a 500 GB file down to 100 GB. With Lena you can go five times smaller, so down to 20 GB, but you can also do it three times faster than simple compression techniques,” adds del Giudice.

ORIGAMI’s market study allowed Enancio to group potential customers into segments and explore how their needs for data compression may differ. One need which stood out was being able to integrate compression into existing processes transparently and without interrupting the workflow.

Demand for sequencing is growing. Over the decade to 2015, genomic data grew at an astonishing rate, doubling once every 7 months according to a study in ‘PLOS Biology’, and this rate of growth is expected to accelerate.

Enancio found many data users have not yet reached the tipping point where data flows have become unmanageable. “The volume of data is not yet causing a problem for everybody, but this will soon come,” observes del Giudice.

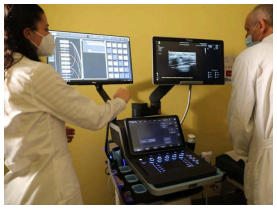
Keywords

ORIGAMI, data compression, compression, genomic data, human genome, sequencing, high throughput sequencing, personalised medicine

Discover other articles in the same domain of application

Explainable AI for personalised therapy in metastatic colorectal cancer

Are there really bacteria in the womb?



Catching up with SOLUS: A new diagnostic approach for breast cancer

Shedding more light on the human brain

Project Information

ORIGAMI

Grant agreement ID: 877363

Project website

DOI
10.3030/877363

Project closed

EC signature date
25 July 2019

Start date
1 August 2019

End date
31 January 2020

Funded under
INDUSTRIAL LEADERSHIP - Innovation In SMEs

Total cost
€ 71 429,00

EU contribution
€ 50 000,00

Coordinated by
ENANCIO
France

Last update: 14 July 2020

Permalink: https://cordis.europa.eu/article/id/421514-new-technique-helps-to-tame-the-deluge-of-genomic-data