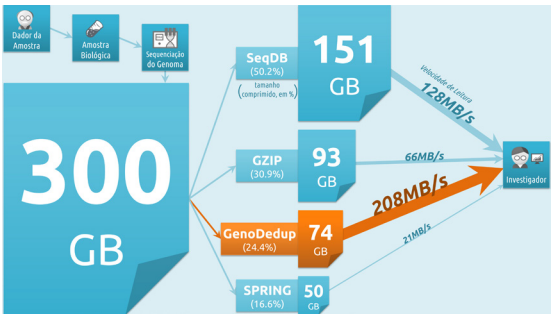
Ciências ULisboa
Faculdade de Ciências da Universidade de Lisboa

Comunicado de Imprensa

4 - agosto - 2020

Tecnologia portuguesa permite acelerar e reduzir custos de estudos na área da saúde



Maior rapidez de leitura e economizar 75% do espaço de armazenamento em dados da sequenciação de genomas humanos é a inovadora solução dos investigadores **Vinicius Vielmo Cogo** e **Alysson Neves Bessani** (**LASIGE - Faculdade de Ciências da ULisboa**) e **João Tiago Paulo** do Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência - (**INESC TEC**) e **Universidade do Minho**.

Combina uma nova técnica de deduplicação de dados baseado em semelhanças e padrões encontrados nos ficheiros de sequenciação de genomas humanos e uma codificação das alterações para a recuperação desses dados.

Substitui-se, assim, a descrição completa dos dados genómicos sequenciados por pequenos apontadores descrevendo-se, apenas, as alterações necessárias para a recuperação dos dados originais, reduzindo proporcionalmente o espaço e o custo de armazenamento.

Esta solução foi publicada na **revista IEEE Transactions on Computers**, uma das mais reconhecidas revistas científicas da área de informática no mundo, no passado dia 14 de maio de 2020, a versão dos autores pode ser consultada em: **artigo principal (main)** e **material suplementar (supp)**.

E por que é relevante? Porque permite aos hospitais e biobancos economizar no armazenamento de dados, ao mesmo tempo que permite que investigadores leiam esses dados de forma mais rápida. Os biobancos e os hospitais são responsáveis por guardar e distribuir milhões de amostras biológicas humanas para investigadores de todo o mundo e, atualmente, estão sob pressão para, também, armazenar os dados genómicos sequenciados a partir destas amostras.

O conhecimento técnico e os orçamentos limitados para a criação de infraestruturas apropriadas para o armazenamento eficiente destes dados instiga a procura de novas soluções que equilibrem a economia de espaço de armazenamento e a velocidade de recuperação/ leitura destes dados.

A aplicação em infraestruturas (p. ex., 1000 Genomes Projects), que já utilizam algoritmos de compressão genéricos nestes dados (p. ex., GZIP), beneficiam de uma redução adicional do custo e espaço de armazenamento (na ordem dos 22%), e permitem que os investigadores acedam a estes dados de forma mais eficiente (5 vezes mais rápido).

E para o futuro? Os investigadores pretendem disponibilizar a solução em código aberto e melhorar os resultados obtidos através de estudos mais aprofundados sobre os padrões e semelhanças entre genomas humanos sequenciados. Este método será também adaptado na sequenciação de genomas de outras espécies, para outras máquinas de sequenciação e outras representações de dados relacionadas.

Informações:
(LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal) Vinicius Vielmo Cogo - *Email*: **vielmo@lasige.di.fc.ul.pt** / Alysson Neves Bessani - *Email*: **anbessani@fc.ul.pt**
(HASLab—High-Assurance Software Lab, INESC TEC & U. Minho, Portugal) João Tiago Paulo - *Email*: **jtpaulo@inesctec.pt**

Ciências ULisboa
LASIGE driven by excellence
INESCTEC
1985 TRINTA E CINCO ANOS 2020