Mark Meehan

Master in Informatics and Computing Engineering
Reproducible and Privacy-preserving Analyses for Next-Generation Sequencing data
Mark Timothy Vasconcelos Meehan
—

Abstract

The analysis of health research data has been growing in relevance and has proven to be very useful across many fields of scientific study. Some highly complex data analysis methodologies, named High-Throughput Technologies (HTT), have been developed to process extremely extensive volumes of patient data. One of the most relevant and standard HTT is Next Generation Sequencing (NGS). Through NGS, researchers can perform very detailed introspections across various biological applications, including sequencing the entire human genome.

To employ such data analysis technologies, researchers resort to bioinformatics pipelines, which are computational pipelines consisting of algorithms capable of extensively processing large datasets of sequencing data. Such pipelines usually require vast volumes of patient data, which is very sensitive and requires special safeguards and procedures to be put in place. Therefore, these data must be analysed in a privacy-preserving manner, which proves to be especially difficult when dealing with extensive datasets that may require execution across a cluster of machines.

A crucial property of bioinformatics pipelines is their reproducibility since they must constantly output relevant results across different builds. Furthermore, reproducibility empowers collaboration since researchers can share bioinformatics pipelines while achieving consistent results. At the moment, this is one of the principal liabilities that bioinformatics pipelines showcase.

This work focuses on studying and improving current frameworks for developing and analysing reproducible bioinformatics pipelines. An actual bioinformatics pipeline will be implemented in DolphinNext, which is considered the most advanced, complete and portable bioinformatics pipeline execution and distribution platform. A set of reproducibility features of the platform are evaluated, including execution output consistency, configuration overhead, portability and shareability. Besides that, different versions of a bioinformatics pipeline are implemented by changing execution-specific parameters. By analysing the similarity of different outputs, important conclusions are drawn from the impact that small changes in pipeline configuration can have on final output results.

From the carried work, it is clear that DolphinNext is a complete and stable platform that can provide accurate and consistent analysis of bioinformatics pipelines. Creating revisions will small changes in a given pipeline is supported natively and proves to be extremely useful in developing and maturing pipelines. Reproducibility seemed to be, overall, ensured by DolphinNext.

Studying a bioinformatics pipeline’s determinism and consistency is crucial for understanding its reproducibility. Therefore, in future work, comparability tools should be supported natively across different pipeline execution and distribution platforms.

Resumo

A análise dos dados de investigação em saúde tem vindo a crescer em relevância e provou ser muito útil em muitos campos de estudo científico. Algumas metodologias de análise de dados altamente complexas, chamadas High-Throughput Technologies (HTT) - têm sido desenvolvidas para processar volumes

extremamente extensos de dados de pacientes. Uma das HTT mais relevantes e correntes é a Next Generation Sequencing (NGS). Através da NGS, os investigadores podem realizar introspecções muito detalhadas através de várias aplicações biológicas, incluindo o sequenciamento de todo o genoma humano.

Para aplicar tais tecnologias de análise de dados, os investigadores recorrem a bioinformatics pipelines, que são pipelines computacionais constituídas por algoritmos capazes de processar extensivamente grandes conjuntos de dados de sequenciação. Tais pipelines requerem normalmente grandes volumes de dados de pacientes, o que é muito sensível e requer salvaguardas e procedimentos especiais a serem postos em prática. Por conseguinte, estes dados devem ser analisados de forma a preservar a privacidade, o que se revela especialmente difícil quando se trata de conjuntos de dados extensos que podem exigir a execução distribuída através de um conjunto de máquinas.

Uma propriedade crucial das bioinformatics pipelines é a sua reprodutibilidade, uma vez que devem produzir constantemente resultados relevantes ao longo de diferentes execuções. Além disso, a reprodutibilidade permite a colaboração, uma vez que os investigadores podem partilhar bioinformatics pipelines alcançando resultados consistentes. Neste momento, esta é uma das principais responsabilidades que as bioinformatics pipelines apresentam.

Este trabalho concentra-se no estudo e melhoria das estruturas actuais para o desenvolvimento e análise de bioinformatics pipelines reproduzíveis. Uma bioinformatics pipelines real será implementada no DolphinNext, que é considerada a mais avançada, completa e portátil plataforma de execução e distribuição de bioinformatics pipelines. É avaliado um conjunto de características de reprodutibilidade da plataforma, incluindo a consistência da execução, a sobrecarga de configuração, a portabilidade e a capacidade e facilidade de partilha. Além disso, são implementadas diferentes versões de uma bioinformatics pipelines através da alteração de parâmetros específicos de execução. Ao analisar a semelhança dos diferentes outputs, são tiradas importantes conclusões do impacto que pequenas alterações na configuração da pipelines podem ter nos resultados finais.

Do trabalho realizado, fica claro que o DolphinNext é uma plataforma completa e estável que pode fornecer uma análise precisa e consistente das bioinformatics pipelines. A criação de revisões irá apoiar pequenas alterações numa determinada pipeline e revela-se extremamente útil no desenvolvimento e na maturação de pipelines. A reprodutibilidade parece ser, de um modo geral, assegurada pelo DolphinNext.

O estudo do determinismo e consistência de uma bioinformatics pipelines é crucial para compreender a sua reprodutibilidade. Portanto, no trabalho futuro, diferentes ferramentas de comparabilidade devem ser implementadas nativamente através de diferentes plataformas de execução e distribuição de *pipelines*.

Jury

- Chair: Ademar Aguiar
- External Examiner: Gur Yaari
- Supervisor: João Correia Lopes
- Date: 20/7/2022