ASSIS, Enock Cavalcante de; CUTIGI, Jorge Francisco. Investigação da organização dos dados de um grande repositório de mutações genéticas em câncer. In: WORKSHOP DE INOVAÇÃO, PESQUISA, ENSINO E EXTENSÃO, 2., 2016, São Carlos, SP. **Anais**... São Carlos, SP: IFSP, 2016. p. 112-115. ISSN 2525-9377.

INVESTIGAÇÃO DA ORGANIZAÇÃO DOS DADOS DE UM GRANDE REPOSITÓRIO DE MUTAÇÕES GENÉTICAS EM CÂNCER

ENOCK CAVALCANTE DE ASSIS[1], JORGE FRANCISCO CUTIGI[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, São Carlos, Brasil.

RESUMO: Novas tecnologias de sequenciamento genômico estão revolucionando o estudo de doenças causadas por falhas no material genético, como o câncer. Essa revolução advém do fato da redução do custo e aumento da velocidade no sequenciamento, o que tem gerado um aumento significativo de dados genéticos nos últimos anos. Nesse contexto, grandes repositórios públicos de dados genômicos foram criados, os quais são mantidos e atualizados constantemente. Dentre os repositórios criados nos últimos anos destaca-se o TCGA (*The Cancer Genome Atlas)*, que possui como objetivo manter um repositório com alto volume de dados com acesso público para que se possa subsidiar pesquisas na área de câncer. Este trabalho apresenta estudos preliminares do repositório TCGA, com o objetivo de entender e descrever os tipos de dados existentes no repositório assim como investigar as formas de extração desses dados.
PALAVRAS-CHAVE: Biologia computacional. Bioinformática. Câncer. TCGA.

ABSTRACT: New genomic sequencing technologies are revolutionizing the studies of disease caused by genetic material failure, such as cancer. This revolution is due to the cost reduction and the increasing sequencing speed, that has brought a significantly increase of genetic data in the last years. In this context, great public genomic data repository has been created, which are maintained and constantly actualized. One of these repositories is TCGA (The Cancer Genomic Atlas), whose objective is to maintain a repository with a high volume of data accessible to the public in order to allow research in the cancer area. This project aims to study TCGA to understand and describe repository existing data types, as well as investigating the way of extracting such data.
KEYWORDS: Computational biology. Bioinformatics. Cancer. TCGA

INTRODUÇÃO

O câncer é uma doença genética que tem desafiado a medicina no que diz respeito à sua prevenção, cura, e quem sabe, a sua erradicação. O câncer origina-se a partir de mutações genéticas específicas conhecidas na comunidade como *driver mutations*, já que nem todas mutações que ocorrem podem ser responsabilizadas pelo seu surgimento. Essas mutações provocam um comportamento anormal de algumas células, que passam a se reproduzir de forma desordenada deixando de responder a sinais que lhes são transmitidos por outras células para que interrompam seus processos de reprodução, produzindo tumores.

Identificar e estudar o que provoca tais mutações tem sido um constante desafio da ciência. Uma área científica promissora para trabalhar nesse desafio é a biologia computacional ou a bioinformática. Criar métodos computacionais que sejam capazes de identificar *driver mutations* exige a existência e disponibilidade de um alto volume de dados,

o que tem sido levado a efeito pelo TCGA (*The Cancer Genome Atlas*), que se trata de uma base de dados com informações genéticas de vários tipos de câncer.

Embora o TCGA seja uma base de dados, o entendimento da organização dos dados e os tipos de dados disponíveis não é uma tarefa trivial, principalmente para cientistas que não possuem conhecimento substancial na área biológica. Nesse contexto, o objetivo deste trabalho é apresentar os resultados preliminares de um projeto que busca estudar os dados do TCGA, investigando como estão disponibilizados e como acessá-los, fornecendo subsídios para que os profissionais da área da computação sejam capazes de acessar as informações com maior facilidade, podendo contribuir nessa área de pesquisa.

MATERIAL E MÉTODOS

Os materiais que foram e estão sendo utilizados na condução deste projeto são artigos científicos e periódicos da área de computação e saúde, que viabilizam o levantamento bibliográfico a respeito do TCGA. Aliado a isso, estão sendo estudados os manuais do TCGA e a organização dos dados na sua estrutura.

O projeto é conduzido considerando as três fases descritas abaixo. Entretanto, este trabalho apresenta os resultados preliminares obtidos na execução parcial da Fase 1 e na execução inicial da Fase 2.

- Fase 1: Levantamento bibliográfico de trabalhos relacionados ao TCGA e manuais de utilização. Resultado esperado: Entendimento da proposta do TCGA, funcionamento e aplicações.
- Fase 2: Investigação dos tipos dados que são armazenados no TCGA e análise e execução de mecanismos para extração desses dados. Resultado esperado: Descrição dos tipos de dados providos pelo TCGA, assim como sua organização e mecanismos para extração.
- Fase 3: Análise dos dados extraídos e execução em métodos computacionais. Resultado esperado: Descrição dos dados de saída e resultado de execução em determinado método computacional.

RESULTADOS E DISCUSSÃO

Dos resultados obtidos até este momento, destaca-se: o levantamento bibliográfico inicial, com o objetivo de entender o TCGA; o entendimento da nomenclatura atribuída a cada conjunto de dados contidos no TCGA; e a confecção de um glossário com termos técnicos referente a bioinformática aplicada ao câncer. Os resultados são descritos nos parágrafos seguintes.

No levantamento bibliográfico inicial foram coletados trabalhos científicos que utilizavam dados do TCGA como base, tal como apresentado por Weinstein (2013). Além disso, constatou-se que o TCGA é resultado do esforço de uma parceria entre o NCI (*National Cancer Institute*) e o NHGRI (*National Human Genome Research Institute*), criado com a finalidade de receber amostras de tecidos afetados e sadios de pacientes voluntários e disponibilizar os dados para a comunidade científica. Seu principal objetivo é prover um fácil acesso às informações reais a respeito de mutações genéticas que podem levar ao surgimento de cada tipo de câncer. Esse volume grande de dados pode ajudar pesquisadores a realizar estudo sobre prevenção, diagnóstico e tratamento de pacientes com câncer, além da criação de métodos computacionais que usem esses dados, conforme os métodos revisados por Cheng (2015) e Raphael (2014). Os dados do TCGA podem ser extraídos por meio de um *site* na *web*[1].

[1] https://tcga-data.nci.nih.gov/tcga/

Os dados do TCGA são armazenados de acordo com uma nomenclatura específica. Buscando entender a metodologia adotada pelo TCGA no que diz respeito à sua principal atividade, observa-se que cada tipo de câncer recebe uma sigla que o identifica e que as amostras de tecidos afetados por um determinado tipo de câncer são entregues a um grupo de trabalho que se dedica ao estudo daquele tipo específico. Exemplo: [BRCA] (*Breast invasive carcinoma*) se trata de uma nomenclatura usada para identificar o câncer de mama. Para cada amostra de tecido recebida de doadores voluntários é criado um código único que relaciona os dados e o local onde a amostra está armazenada, utilizando os seguintes dados: origem (doador); pesquisador responsável pelo desenvolvimento do projeto; tipo da amostra; ordem em que foi armazenada (A-Z); porções dela retiradas para estudo, entre outras informações. Por exemplo, no código TCGA-02-0001-01C-01D-0182-01, TCGA indica onde o projeto está sendo desenvolvido; 02 indica a origem do tecido; 0001 indica quem participa do estudo; 01 indica o tipo de tecido; C indica a ordem ocupada pelo frasco que contém a amostra; 01 indica a primeira porção da amostra (que pode ir de 01 a 99); D indica que o componente é uma amostra de DNA; 0182 indica a ordem em que se encontra entre outras amostras; e, finalmente, 01 indica o centro de sequenciamento ou caracterização que receberá a alíquota para análise.

Um glossário, em português, com termos técnicos referentes a bioinformática e câncer também tem sido elaborado no decorrer do projeto. O glossário tem como propósito facilitar o entendimento dos profissionais da área da computação não familiarizados aos termos utilizados na área de saúde. Espera-se selecionar os termos mais relevantes aos profissionais leigos, com o objetivo de facilitar a entrada de profissionais da computação que não são familiarizados com termos biológicos. Por exemplo, o termo "*driver mutation*", supracitado, possui uma definição difícil de ser entendida com a tradução literal e, por ser um termo comum na área biológica, há dificuldade em obter a definição nos artigos científicos. Ao final do projeto, o glossário será disponibilizado para acesso público.

CONCLUSÕES

Embora o TCGA possa ser usado publicamente, os dados nele existentes são de difícil acesso e compreensão, principalmente por pesquisadores de outras áreas, diferente da genética, pois eles envolvem muitos termos técnicos e específicos. Com isso, é válido o esforço para o entendimento do conteúdo e do mecanismo de extração desses dados para uso posterior. A primeira tentativa para se diminuir essa dificuldade foi a identificação por meio de um código único, o qual foi entendido e permite reunir as informações de origem e armazenamento dos dados.

A criação de um glossário com termos biológicos ligados ao câncer vem, de certa forma, facilitar o entendimento de determinados termos inerentes à área biológica por novos pesquisadores da área de computação, diminuindo o tempo que seria despendido nessa atividade. Os autores acreditam que a disponibilização do referido glossário é de grande utilidade, sendo um resultado importante a ser obtido no projeto.

Como continuação do projeto, pretende-se entender, além da organização dos dados, a maneira como se pode extraí-los, ou seja, quais os mecanismos existentes no TCGA para que os dados possam ser obtidos e utilizados por pesquisadores da área de computação. Espera-se que um roteiro de extração desses dados seja elaborado e disponibilizado em português, assim como o glossário.

REFERÊNCIAS

CHENG, Feixiong; ZHAO, Junfei; ZHAO, Zhongming. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. **Briefings in Bioinformatics**, 2015.

RAPHAEL, Benjamin J.; DOBSON, Jason R.; OESPER, Layla; VANDIN, Fabio. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. **Genome Med.**, 6:5, 2014.

WEINSTEIN, John N.; COLLISSON, Eric A; MILLS, Gordon B.; SHAW, Kenna R. M.; OZENBERGER, Brad A, ELLROTT, Kyle; SHMULEVICH, Ilya; SANDER, Chris; STUART, Joshua M. The Cancer Genome Atlas Pan-Cancer analysis project. **Nat Genet.,** 45:1113–20, 2013.