Análise de sequenciamento genético em *proband* para identificar as mutações genéticas e divergências encontradas

Jeferson da Silva Grigolo[1], **Sylvio André Garcia Vieira**[1]

[1]Sistema de Informação – Universidade Franciscana - UFN
Caixa Postal 151 – 97010-032 – Santa Maria – RS – Brasil

jeferson.grigolo@ufn.edu.br, sylvio@ufn.edu.br

Abstract. *Colorectal cancer (CRC) is the fourth crucial reason for death worldwide, the evolution of which is due to the accumulation of genetic mutations. The objective of this work is to carry out sequence analysis and develop software to find genetic mutations, identify the positions and respective cracks of the nitrogenous bases that form an amino acid. To achieve this purpose, the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology was used, the Python programming language using the Django framework and with the help of Biopython. The product of this work is a tool that can assist health professionals in research and detection of CRC mutations.*

Resumo. *O câncer colorretal (CRC) é a quarta crucial razão de morte no mundo inteiro, a evolução dele é devido ao acúmulo de mutações genéticas. O objetivo deste trabalho define-se em efetuar a análise de sequência e desenvolver um software para encontrar as mutações genéticas, identificar as posições e as respectivas trincas das bases nitrogenadas que formam um aminoácido. Para conseguir essa finalidade foi utilizada a metodologia Cross Industry Standard Process for Data Mining (CRISP-DM), a linguagem de programação Python com o emprego do framework Django e com o auxílio do Biopython. O produto deste trabalho, é uma ferramenta que pode auxiliar os profissionais da área da saúde, nas pesquisas e detecção das mutações do CRC.*

1. Introdução

Os estudos na área de saúde vêm evoluindo com novos meios, recursos e sistemas, desenvolvidos pela Tecnologia de Informação e Comunicação (TIC) na forma de auxiliar no diagnóstico antecipado e oferecer dados relevantes para os profissionais da área. Com o auxílio da TIC é possível armazenar, processar e comunicar a informação [Araújo et al., 2019].

Análise de sequência é um método aplicado para definir a ordem linear dos nucleotídeos em uma parte do DNA. A análise de sequência pode contribuir na detecção de mutações genéticas em uma pessoa [Paula 2018].

Análise de sequenciamento genético tem potencial de colaborar com os profissionais de saúde como: localizar mutações genéticas, diagnóstico na medicina para encontrar doenças [Castro 2015].

Dentre os possíveis diagnósticos das doenças cancerígenas, há o câncer colorretal (CRC). A evolução do CRC é um processo passível de ser sequenciado, que compreende

o acúmulo evolutivo das mutações, que procedem da ativação de oncogenes e estagnação de genes supressores de tumor. Essa evolução cancerígena transcorre de *probands* (conjunto de dados) normais para adenoma, que caracteriza-se por ser um primeiro estágio tumoral, em que o organismo ainda possui condições de combatê-lo, até chegar no estágio de carcinoma, que já é um estágio avançado da doença necessitando de adição de medicamentos para ser combatido.

De acordo com o Instituto Nacional do Câncer (INCA) INCA (2019) ligado ao Ministério da Saúde, possui estimativa de CRC, para cada ano do triênio de 2020-2022, o surgimento de 20.520 novos casos em homens e 20.470 em mulheres. Segundo Gandra (2020) na reportagem realizada com o cirurgião Marcus Valadão, do INCA, especialista em CRC, essa doença é considerada silenciosa, em estágios iniciais não exibe sintomas. Quando detectado precocemente o CRC, tem uma possibilidade de cura entre 90% e 95%.

Baseando-se no alto número estimado de ocorrências do CRC, e com possibilidade de cura acima de 90%, se detectado precocemente; avaliou-se o desenvolvimento de um software que execute a análise de sequência para detectar mutações genéticas e divergências do CRC no *proband*. Esse sistema faz conexão com o servidor NCBIWWW Blast, que tem a finalidade de encontrar e salvar dados com alinhamentos na base de dados do *National Center for Biotechnology Information* (NCBI). Após esse processo, o sistema executa a análise de sequência nos dados encontrados, procurando as mutações e divergências com as respectivas posições e trincas das bases nitrogenadas, que formam um aminoácido, para auxiliar profissionais da área da Saúde.

Como sugestões para trabalhos futuros, há possibilidade de otimizar a consulta e o alinhamento de sequência junto ao servidor NCBIWWW Blast, por intermédio de Web Services e/ou API's e aperfeiçoar o banco de dados do sistema, passando do DB sqlite3 para MySQL ou similares.

1.1. Objetivos Geral

Desenvolver um software para encontrar as mutações genéticas, identificar as posições e as respectivas trincas das bases nitrogenadas que formam um aminoácido, por meio da análise do sequenciamento genético.

1.2. Objetivos Específicos

- Desenvolver um software com linguagem de programação Python e o framework web Django, com auxílio da biblioteca Biopython utilizando os módulos Blast, Seq e SeqIO
- Realizar análises de sequenciamento genético nas bases nitrogenadas de *proband* para encontrar as mutações e divergências encontradas com as respectivas posições e as trincas das bases nitrogenadas que formam um aminoácido.

2. Referencial Teórico

Nesta seção serão abordados os tópicos: gene, ácidos nucleicos, câncer colorretal, linguagem de programação, Biopython, Django, análise de sequência, CRISP-DM, Uml e banco de dados biológicos.

2.1. Gene

Gene é um encadeamento de nucleotídeos de DNA, encarregado pelos atributos herdados geneticamente. Determinados genes atuam com orientações para produzir moléculas denominadas proteínas. Nos seres humanos os genes diferenciam-se no tamanho das bases de DNA, podendo conter mais de 2 milhões. O Projeto Genoma Humano estimou que os humanos têm entre 20.000 e 25.000 genes [Medicine 2020a].

Dentro da literatura da genética atual, o gene é uma sequência de nucleotídeos de DNA que é capaz de efetuar uma transcrição para o RNA. O gene é um fragmento do cromossomo que se refere a um código característico, e este possui informação necessária para gerar uma deliberada proteína ou coordenar uma particularidade, por exemplo a cor dos olhos [Marques 2018].

As pessoas possuem duas cópias de cada gene recebidas dos seus pais. A maior parte dos genes é equivalente em todos os humanos, entretanto menos de 1% do total é dissemelhante entre os humanos. Os alelos são moldes do mesmo gene com poucas diferenças no sequenciamento das bases de DNA, que resultam em aptidões físicas particulares de cada ser humano [Marques 2018].

2.2. Ácidos nucleicos

Os ácidos nucleicos são caracterizados por macromoléculas grandes, constituídas por unidades monoméricas, classificado como nucleotídeos. Há duas categorias de ácidos nucleicos: Ácido Desoxirribonucleico (DNA) e Ácido Ribonucleico (RNA). Eles são encarregados de codificar e interpretar os dados que especificam a síntese de proteínas dos seres humanos [Santos 2020].

Os ácidos nucleicos são constituídos de nucleotídeos que são moléculas formadas pelas junções de três elementos: Ácido fosfórico, pentoses e base nitrogenada. O ácido desoxirribonucleico (DNA) é a molécula que acumula dados genéticos. É construído por três categorias de nucleotídeos e quatro tipos de bases nitrogenadas, das classes purinas e pirimidinas: Purinas primárias são adenina (A) e guanina (G) e pirimidinas comuns são a timina (T) e a citosina (C), que irá produzir moléculas de DNA diferentes de acordo com a quantidade dos nucleotídeos e o sequenciamento [Nelson and Cox 2014].

No processamento da transição, a informação salva do DNA é replicado no RNA através da atuação da enzima RNA polimerase que efetua a abertura e exibição das sequências de nucleotídeos que será transcrito utilizando apenas uma das fitas de DNA para realizar a síntese de proteínas [Lodish et al., 2014].

2.3. Câncer Colorretal

Segundo o INCA (2020) o câncer é originado de alterações da estrutura genética do DNA das células, mais conhecidas como mutações genéticas. Os genes possuem instruções que devem executar, como crescimento, divisão, morte celular. Ao ocorrer mutações genéticas, as instruções dos genes podem ser alteradas, que poderá dar início ao surgimento do câncer. Em muitos casos ocorre: multiplicação descontrolada, desprender-se do adenoma e deslocar-se para regiões próximas, perda das instruções originais.

O CRC atinge tanto homens quanto mulheres, é o terceiro e segundo câncer mais frequente respectivamente. Com cerca de 1,4 milhão de novos casos durante um ano. Tornou-se a quarta crucial razão de morte no mundo inteiro. Os médicos utilizam a taxa

de sobrevida como uma forma padrão para debater o prognóstico de um paciente, que se refere à porcentagem de pacientes que vivem pelo menos 5 anos após o diagnóstico da doença. O CRC possui uma taxa de sobrevida em cinco anos abaixo de 65%. A evolução do CRC pode ser estimulada pelas ligações de vários fatores genéticos e ambientais de modo complexo, como, NCI (instabilidade cromossômica), MSI (instabilidade microssatélites), fenótipo metilador[1] da CIMP (ilha 5'-C-fosfato-G-3' - CpG)[2], e hipermetilação global do DNA. O CIMP é identificado pois atinge as regiões promotoras do gene, ao metilar diversas ilhas de CpG. Ao silenciar a transcrição dos genes supressores tumorais o CIMP é classificado como uma das cruciais vias tumorigênicas no CRC [Zhang et al., 2020].

O CRC é um adenoma que afeta o intestino grosso, compreendido pelo sigmoide, o reto e os cólons: ascendente, transverso e descendente. A maior parte das neoplasias que afeta a área do colorretal é categorizado como adenocarcinoma, tipo de neoplasia maligna que se origina nas células glandulares produtoras de muco que envolvem o cólon e o reto [Cordeiro 2019].

O CRC é identificado pelo desenvolvimento incontrolado de células epiteliais do cólon e reto, relacionado com específicas mutações genéticas que resultará em formação de adenoma e carcinoma. Dentre as mutações relacionadas ao CRC destacam-se mutações que inativam os genes supressores tumorais APC e P53 e ativam oncogenes RAS, BRAF e PI3K [Cordeiro 2019].

Segundo Guedes et al. (2019) o CRC é a neoplasia mais comum nos EUA, no Brasil ocupa a quarta colocação, com mais relevância nos estados de São Paulo e Rio de Janeiro. Em grande parte dos CRC progride do início de adenomas preestabelecido. A ativação de genes relacionados com o aparecimento e crescimento de tumores e a inativação de genes supressores de tumor são a estrutura molecular para a manifestação do carcinoma. Mutações no gene TP53 que codifica a proteína p53, está vinculada no progresso de diversos carcinomas humanos. Quando ocorre o TP53 descontrolado.

2.4. Linguagem de Programação

Para desenvolver o software proposto, foi utilizado Python, que é uma linguagem de programação de alto nível, interpretada, orientada a objetos com semântica dinâmica, possui uma grande performance e dispõe de vasta quantidade de bibliotecas [Python 2020].

2.5. Biopython

Biopython é um projeto internacional da associação dos desenvolvedores de ferramentas para a biologia molecular computacional. Deste projeto surgiu a biblioteca Biopython escrita em Python, que possui um conjunto de ferramentas que está à disposição gratuitamente para utilização na bioinformática. Sua finalidade é proporcionar uma melhor usabilidade do Python, com módulos e classes reutilizáveis e com alta qualidade.

---

[1] A metilação é uma reação química que consiste na adição de um grupo metila (-CH3) a uma molécula, com o caso particular quando um átomo de hidrogênio é substituído por um grupo metil.

[2] Ilha 5'-C-fosfato-G-3' abreviatura de CpG são regiões do DNA onde um nucleotídeo de citosina é seguido por um nucleotídeo de guanina na sequência linear de bases ao longo de sua direção 5 '→ 3'. Os locais CpG ocorrem com alta frequência em regiões genômicas chamadas ilhas.

O site Biopython[3] oferece recursos para módulos e scripts e links para desenvolvedores de software baseado em Python. As capacidades do Biopython abrangem analisadores de diversos formatos de arquivos para bioinformática e fornecem suporte para várias estruturas de dados como: Fasta[4], GenBank[5], Blast[6], Clustalw, NCBI, ExPASy, SwissProt, Unigene [Biopython 2020a].

2.6. Django

Django é um framework de web *server-side*, desenvolvido em Python, é um dos frameworks mais conhecidos entre os programadores. É gratuito e com código aberto, simplificador ao desenvolvimento de várias aplicações. Django aplica o padrão de projeto *Model, Template and View* (MTV). O *Model* é onde ocorre a comunicação com o banco de dados. O *Template* é realizado a renderização para uma interação do site em django com o usuário. *View* é o nível responsável pelas regras de negócios dados, e as requisições dos usuários, por intermédio de aplicações como o roteamento de URLs[7] [Cardoso and Bispo 2019].

2.7. Análise de Sequência

Análise de sequência é um procedimento para encontrar a sequência correta para a formação do aminoácido. É possível identificar se a composição da trinca de bases nitrogenadas formou um aminoácido válido [Biometrix 2018].

O sequenciamento genético possibilita detectar uma gama inteira de variações genéticas comuns e outras diferentes do normal. Proporciona auxiliar nas doenças raras e utilidades clínicas, que viabiliza obter detecção, caracterização e previsibilidade. [Lappalainen et al., 2019].

2.8. CRISP-DM

Uma das metodologias mais utilizadas em projetos que abrangem a mineração de dados é a *Cross Industry Standard Process for Data Mining* (CRISP-DM). Ela foi desenvolvida inicialmente em 1996, e em 1997 foi incorporada em um projeto da União Europeia na esfera da iniciativa do aporte ESPIRIT. Esse projeto foi administrado por seis empresas [Junior 2014]. Segundo Andrade et al. (2020) CRISP-DM é amplamente utilizada em projetos que abrangem a mineração de dados. Possui adaptabilidade para diversas finalidades, pois possui fases flexíveis. Segundo Caldas (2019) a quantidade de dados coletados em todas as áreas, está aumentando a cada dia que passa, e os processamentos e análises dos dados em muitas das vezes é lento e demorado, delongando para que sejam utilizados de forma eficiente e inteligente. O CRISP-DM é constituída de um ciclo de seis fases. Que inicia com entendimento de negócio até avaliação e aplicação, com modelo iterativo nas fases, possibilita a regressão e progressão, seguindo os resultados obtidos em cada fase conforme demonstrado na Figura 1.

---

[3] Site Biopython https://biopython.org/.
[4] FASTA é um formato baseado em texto para representar sequencias de nucleotídeos.
[5] GenBank é um banco de dados de anotações de sequências de nucleotídeos.
[6] Blast é uma ferramenta de pesquisa de alinhamento e similaridade local.
[7] URLs é uma sigla de *Uniform Resource Locator* que é o endereço eletrônico.

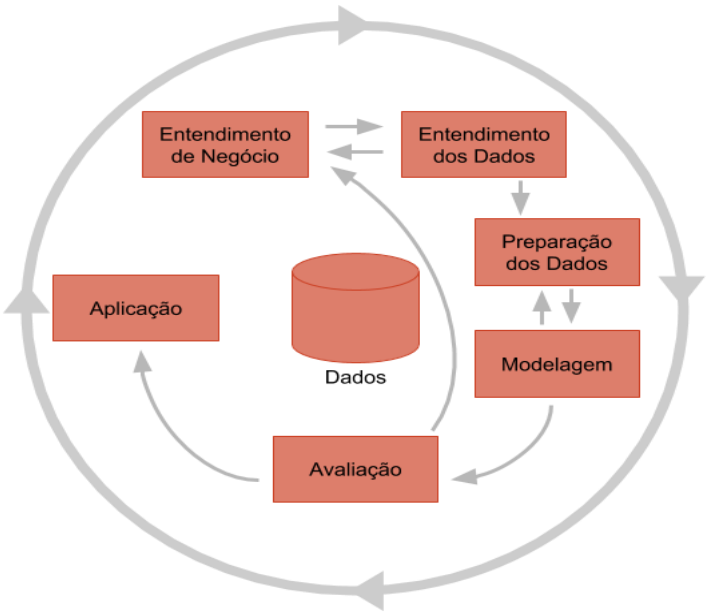


Figura 1. Fases do CRISP-DM [Moura 2019].

2.9. UML

A Linguagem Unificada de Modelagem (UML) é uma linguagem padrão para a elaboração da estrutura de projetos de Software desenvolvido. UML pode ser utilizado para: visualização, especificação, levantamento dos requisitos e a documentação de elementos que façam uso de sistemas complexos de software. A UML abrange uma série de notações para a elaboração de diagramas, representando diferentes aspectos de um software, dentre eles estão os diagramas de: atividades, casos de uso, diagrama de sequência, diagrama de classes entre outros [Booch, Rumbaugh and Jacobson, 2005].

2.10. Banco de dados Biológicos

A quantidade de dados biológicos disponível, vem expandindo de forma exponencial, decorrente dos atuais avanços, nas ciências biológicas e na computação, com modernizações do sequenciamento de próxima geração e a computação de alto desempenho [Guedes et al., 2020].

O Senador Claude Pepper admitiu a relevância de métodos informatizados de processamento de informações para a realização de pesquisas biomédicas, que instituiu NCBI em 1988. O NCBI foi incumbido de realizar diversas atividades, que viessem desenvolver novas tecnologias de informação para contribuir na compreensão de processos moleculares e genéticos, que controlam a saúde e as doenças. Algumas das atividades são: desenvolvimento de sistemas automatizados para armazenamento e análises dos conhecimentos sobre biologia molecular; bioquímica e genética; possibilitar o uso de bancos de dados e software pelas associações médicas e de pesquisas; coordenar empenhos para coletar informações de biotecnologia, tanto nacional como internacionalmente; e realizar o gerenciamento do desenvolvimento, distribuição, apoio e a coordenação ao acesso a uma diversidade de banco de dados e software para as organizações científicas e médicas [Medicine 2020b].

Dentre diversos bancos de dados biológicos públicos, estão o PUBMED[8] que é gratuito e o mais conhecido na área da saúde, GEO[9] é um repositório público de dados genômicos funcionais, todos mantidos pelos NCBI.

3. Trabalhos Correlatos

Nesta seção serão apresentados trabalhos com características análogos, ou com aplicabilidades relacionadas para o desenvolvimento deste trabalho final de graduação.

3.1 Identification of recurrent and novel mutations by whole-genome sequencing of Colorectal tumors from the Han population in Shanghai, eastern China

Em Teng et al. (2018) foi realizado sequenciamento genético em 10 pacientes com CRC, e efetuou a comparação com *probands* normais, com objetivo de encontrar o perfil de mutação genômica de CRC. Foi realizado comparação desses dados com o projeto TCGA[10], que resultou em alguns genes já mencionados nesse projeto, são eles: TP53, KRAS, FAM47C e MUC7.

3.2 Evaluation of gene-environment interactions for colorectal cancer susceptibility loci using case-only and case-control designs

No estudo realizado por Song et al. (2019) para investigar a interação do gene e ambiente (G x Es), aplicou estudo de caso único de CRC, utilizando 703 casos de CRC e 1406 casos controles. O trabalho utilizou 31 SNPs[11] previamente identificados em locais de vulnerabilidade ao CRC e 13 riscos ambientais apresentados pelo GWAS[12]. O trabalho conclui que diante aos resultados apresentados, existem prováveis influências entre o SNP rs4444235 no cromossomo 14q22.2 e exercício frequente e o SNP rs2423279 no cromossomo 20p12.3 e a utilização contínua de aspirina no CRC.

3.3 Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21

No trabalho de Tenesa et al. (2008), foi realizado uma varredura de associação abrangente (GWAS) de genoma em fases. Foi selecionado os controles e foram pareados por sexo, idade, área de residência. Na fase 1 e 2 foram selecionados 1.012 e 2.057 respectivamente, todos com casos de CRC com início precoce em ambas as fases. Na fase 2 foi utilizada a classificação dos SNPs por estatísticas de testes e selecionados os tops 15.008 para análise. Foi utilizado 2.111 controles, que resultou no selecionamento dos cinco SNPs mais bem qualificados em ambas as fases. Na fase 3 da genotipagem realizada, encontrou três das cincos associações, mais bem classificada dos SNPs e seus cromossomos.

3.4 Considerações sobre os trabalhos correlatos

Os trabalhos correlatos citados ilustram diversos aspectos relacionados ao desenvolvimento do presente trabalho. No trabalho de Teng (2018) foi realizado o sequenciamento genético de pacientes com CRC e efetuado a comparação com *probands*

[8] O PubMed compreende mais de 30 milhões de citações biomédicas, https://pubmed.ncbi.nlm.nih.gov/.
[9] https://www.ncbi.nlm.nih.gov/geo/.
[10] Projeto TCGA é um projeto que objetiva cadastrar e descobrir as principais alterações genômicas. .
[11] SNPs é o polimorfismo de nucleotídeo simples, que ocorre uma variação na sequência de DNA.
[12] GWAS é definido como estudo de associação genômica ampla, detecta variações genéticas.

normais para encontrar o perfil de mutação genômica de CRC. No trabalho de Song (2019) foi realizado a interação de gene e ambiente, utilizando os *probands* com CRC e normais. O trabalho de Tenesa (2008) foi realizado varredura de associação genômica em fases, utilizando casos controles e casos com CRC. Estes trabalhos contribuíram com os resultados identificados e a busca por associação nos dados, propiciou mais clareza, nas varreduras e análise dos dados dos *probands* que foram sequenciados.

4. Metodologia

Para desenvolvimento deste trabalho, foi aplicada uma adaptação da metodologia CRISP-DM. Essa escolha está relacionada à prática conveniente em processos de análise de dados, agregado a uma ampla pesquisa bibliográfica, incluindo artigos, livros e vídeos.

4.1. Entendimento de Negócio

Nessa fase, concentra-se na clareza do propósito que se espera obter com a análise de sequência. Para este trabalho foi aplicado análise de sequenciamento do CRC, para identificar os asterisco e divergências nos resultados obtidos com a utilização da biblioteca Biopython, importando os módulos Blast, Seq e SeqIO. O asterisco ocorre quando uma trinca de bases nitrogenadas ao ser transformada para aminoácido, o Biopython identifica que não é válido, identificando com essa marcação (*). Exemplificando: a trinca de bases nitrogenadas timina, guanina e adenina (TGA) na transformação para aminoácido forma um (*). A divergência nos resultados ocorre quando o alinhamento de sequências do arquivo e o retorno do servidor, não são idênticos. Então, o software armazena e salva a divergência na posição inicial e final, da ocorrência desse alinhamento incorreto.

4.2. Entendimento dos Dados

Nessa segunda fase foi preciso organizar, documentar, entender, distinguir os dados relevantes que foram analisados, para obter a descoberta dos resultados: com asterisco e divergências encontrados ao finalizar todas as fases. Essa fase é classificada como complexa, pois investiga a essência dos dados. É necessário entender os conceitos da biologia molecular, para realizar as análises dos dados, como: composição de uma proteína, processo de duplicação de uma molécula de DNA de dupla hélice e mutações celulares. Os dados do *proband* com CRC estão em bancos de dados públicos, específicos para área da Saúde, como os mantidos pelo NCBI.

4.2.1. Requisitos do Sistema

Para melhor compreensão do trabalho, observando a metodologia é necessário especificar as funcionalidades que atendem aos requisitos do sistema desenvolvido. Foi realizado o levantamento dos Requisitos Funcionais (RF) e dos Requisitos Não Funcionais (RNF). Estão listados nas Tabela 1 e 2.

Tabela 1. Requisitos Funcionais

Requisitos Funcionais		
Funcionalidade /RF	Descrição	Complexidade
1: Efetuar Login	O sistema possui tela de login para acessar o software, o usuário deverá entrar com email e senha.	Média
2: Gerenciar Usuários	O sistema efetua o gerenciamento dos usuários do sistema.	Média

3: Enviar Arquivo	O sistema o envio de arquivo com extensões fasta, que contenham sequências das bases nitrogenadas do *proband*.	Baixa
4: Notificar usuário	O sistema notifica o usuário, em caso de envio de extensão incorreta de envio.	Média
5: Conectar ao Servidor	O sistema conecta ao servidor NCBI Blast, para encontrar alinhamentos entre sequências, e salvar em arquivo txt	Alta
6: Analisar o Sequenciamento Genético	O software realiza análise de sequenciamento genético sobre o arquivo enviado e nos alinhamentos encontrados no servidor NCBI Blast.	Alta
7: Salvar os resultados	O sistema salva os asteriscos encontrados e divergências.	Alta

Tabela 2. Requisitos Não Funcionais

Requisitos Não Funcionais		
Funcionalidade/ RNF	Descrição	Complexidade
1: Software	Uso de um Software Web.	Alta
2: Linguagem	O Software será desenvolvido com a linguagem de programação Python utilizando a biblioteca Biopython e o framework web Django.	Média
3: Disponibilidade	O sistema deve permanecer sempre disponível (online) ao usuário.	Alta
4:Validar arquivo	O sistema deverá validar, o arquivo enviado, caso esteja incorreto, notificar o usuário. Com aviso na tela.	Alta
5: Salvar os Dados	Sistema salva os dados da conexão com servidor e análise de sequência em arquivo formato texto (txt).	Alta

4.2.2. Diagrama UML

Para uma melhor compreensão das funcionalidades e controle das atividades do Software desenvolvido, as Figuras 2.1 e 2.2 exibem o diagrama de atividade que o Software executará. Apresenta o fluxo de controle de uma atividade para outra.

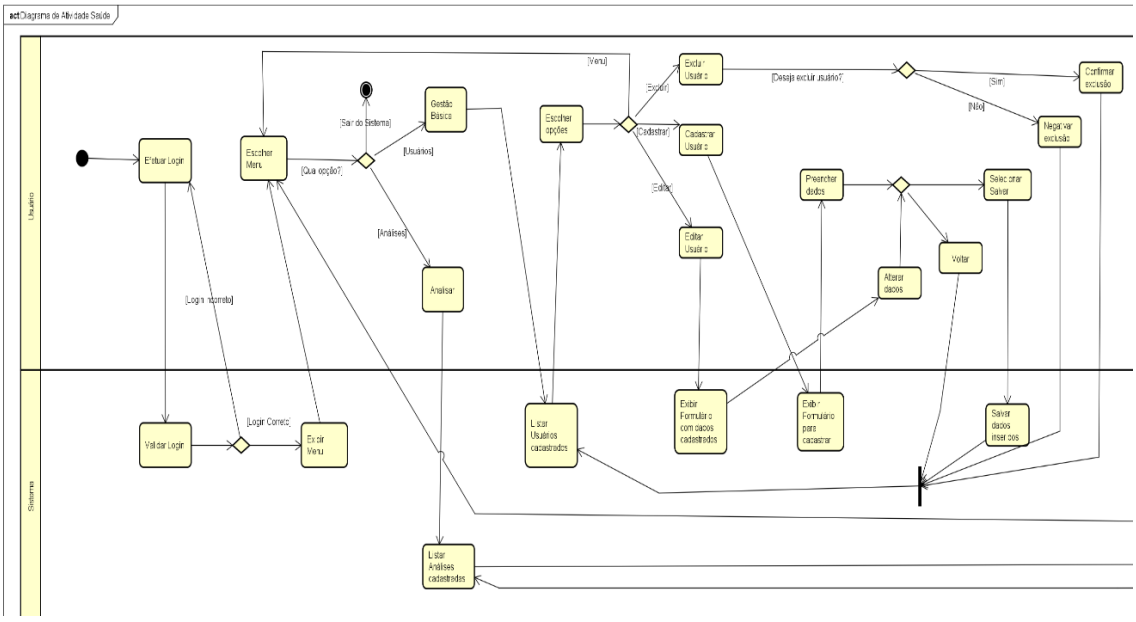


Figura 2.1. Diagrama de atividade [Dos Autores].

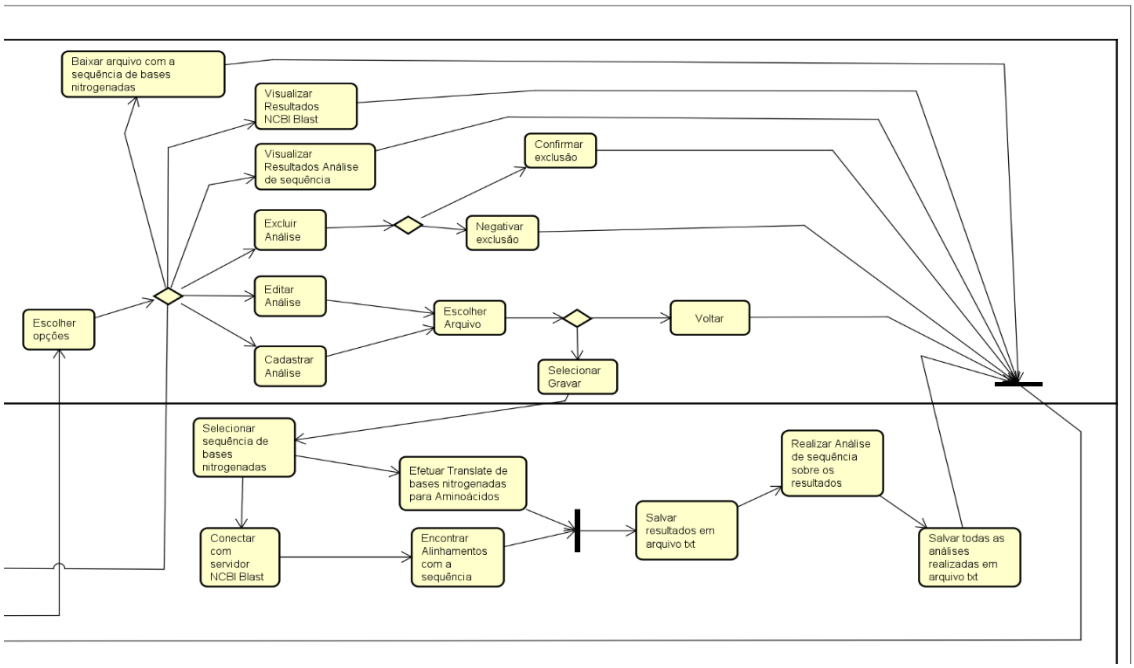


Figura 2.2. Diagrama de Atividade [Dos Autores].

4.2.3. Interface do Usuário

A Figura 3A exibe a interface do login de acesso ao Software, onde o usuário deverá entrar com e-mail e senha para acessar o sistema e a interface Home, conforme Figura 3B. O usuário pode efetuar a gestão básica com o controle dos usuários do sistema e análises, sendo possível visualizar análises ou submeter o arquivo para efetuar as análises de sequenciamento genético.

Figura 3. A- Layout Login e B- Home do Software [Dos Autores].

4.3. Preparação dos Dados

Na preparação dos dados foi necessário selecionar todo o sequenciamento das bases nitrogenadas, pois o arquivo ".fasta" enviado pelo usuário contém informações como: nome, identificação e todo o sequenciamento das bases nitrogenadas, para que o software realize de forma correta as conexões com a biblioteca Biopython.

Nesta fase foi desenvolvido o Software, utilizando a linguagem de programação Python com framework Django, utilizando o banco de dados DB sqlite3 para salvar as

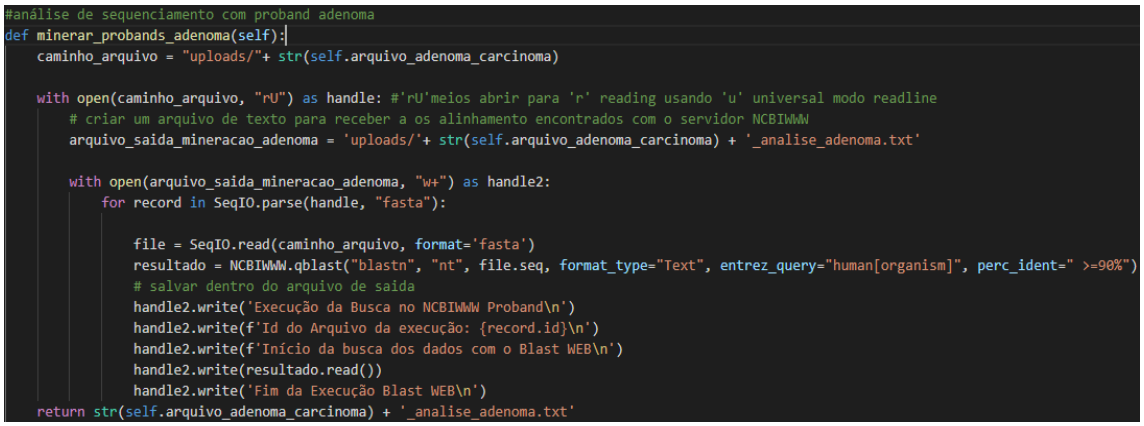
informações do sistema. Foi utilizado a biblioteca Biopython para: efetuar conexão com o servidor NCBI Blast, transformação de bases nitrogenadas em aminoácidos, obter a sequência das bases nitrogenadas no arquivo, realizar o sequenciamento genético, encontrar mutações, localizar divergências e salvar os resultados.

4.4. Modelagem

Nessa fase foi realizada a seleção das importações de pacotes necessários para encontrar os dados do servidor e resultados ao aplicar análise de sequência. Foi utilizado os pacotes Bio.Blast, Bio.Seq e Bio importando SeqIO. Esses 3 pacotes pertencem a biblioteca Biopython.

O software realiza duas etapas após o usuário submeter o arquivo: primeira etapa o software efetua conexão com o 'NCBIWWW.qblast', que é uma funcionalidade para encontrar e retornar os alinhamentos de sequências e salvar essa etapa em um arquivo texto (txt). A segunda etapa, o software aplica análise de sequenciamento no arquivo texto salvo na etapa anterior, buscando encontrar erros e divergências encontrados nas trincas de bases e nas 'Querys' e 'Sbjcts'. O sistema salva em outro arquivo texto todos esses resultados da segunda etapa.

Para realizar as análises de sequências no *proband* do arquivo enviado pelo usuário foi utilizado o pacote Bio importando SeqIO, para encontrar somente as bases nitrogenadas. Após esse processo, foi realizado conexão com o servidor NCBI utilizando o pacote Bio.Blast e a funcionalidade 'NCBIWWW.qblast'. Os parâmetros do método como o programa blastn, compara uma sequência contra uma base de dados de sequência de nucleotídeos. Além de realizar a filtragem de organismo para Homo Sapiens e percentagem de identidade das bases nitrogenadas. Localiza os alinhamentos de sequências com as 'Querys', que são do arquivo enviado pelo usuário e os 'Sbjcts', que são os alinhamentos encontrados pelo servidor. Esses dados são exibidos a cada novo número de arquivo de alinhamentos, encontrado no servidor. A Figura 4 exibe o método desenvolvido para realizar a leitura do arquivo, conexão com servidor e salvar em um arquivo texto “.txt”. Os dados encontrados com a resposta do servidor, encerram a primeira etapa.



```
#análise de sequenciamento com proband adenoma
def minerar_probands_adenoma(self):
    caminho_arquivo = "uploads/"+ str(self.arquivo_adenoma_carcinoma)

    with open(caminho_arquivo, "rU") as handle: #'rU'meios abrir para 'r' reading usando 'u' universal modo readline
        # criar um arquivo de texto para receber a os alinhamento encontrados com o servidor NCBIWWW
        arquivo_saida_mineracao_adenoma = 'uploads/'+ str(self.arquivo_adenoma_carcinoma) + '_analise_adenoma.txt'

        with open(arquivo_saida_mineracao_adenoma, "w+") as handle2:
            for record in SeqIO.parse(handle, "fasta"):

                file = SeqIO.read(caminho_arquivo, format='fasta')
                resultado = NCBIWWW.qblast("blastn", "nt", file.seq, format_type="Text", entrez_query="human[organism]", perc_ident=" >=90%")
                # salvar dentro do arquivo de saida
                handle2.write('Execução da Busca no NCBIWWW Proband\n')
                handle2.write(f'Id do Arquivo da execução: {record.id}\n')
                handle2.write(f'Início da busca dos dados com o Blast WEB\n')
                handle2.write(resultado.read())
                handle2.write('Fim da Execução Blast WEB\n')
    return str(self.arquivo_adenoma_carcinoma) + '_analise_adenoma.txt'
```

Figura 4. Método para conexão e salvar os dados NCBIWWW [Dos Autores]

Na segunda etapa, o software realiza a análise de sequência no arquivo de texto, salvo na primeira etapa. Foi desenvolvido um método aplicando lógica de programação para efetuar as análises de sequência. Nesse método primeiramente é realizado a

transformação de toda a sequência de bases nitrogenadas para aminoácidos utilizando o pacote Bio.Seq e o método Seq.translate. O método junta a cada trinca de bases nitrogenadas e transforma em um aminoácido até o final da sequência. Com objetivo de encontrar os asteriscos (*) nessa transformação.

Após é verificado se a Query da linha contém 60 itens. Ao contemplar essa condição será verificado se o Sbjct do alinhamento com a Query está diferente uma da outra, ou seja, se ocorreu alguma alteração de base ou mutação em uma ou vários itens. O método irá salvar em uma lista: Query e número, início e o fim da mesma. Em seguida, o sistema percorre os aminoácidos transformados das bases, somente entre o início e fim da Query, que está sendo analisado.

Os asteriscos (*) que aparecem, ocorrem devido a troca de qualquer base que está sendo alterado, ficando comprometido o aminoácido gerado, dessa trinca de bases. Após esse procedimento o sistema adiciona em uma lista a sequência do Sbjct e os asteriscos encontrados com a posição na sequência de bases e a trinca que formam esse asterisco. O sistema salva em uma lista as possíveis combinações de trinca de bases e o aminoácido gerado com a nova trinca sugerida. Em seguida, o processo do software procura todas as divergências encontradas na Query e o Sbjct que está analisado. O sistema armazena em uma outra lista o início e fim dessas divergências. É percorrido toda a lista para encontrar as trincas com divergências e salvar essas informações obtidas na análise de sequência. A Figura 5, exibe a resposta do servidor NCBIWWW com as Querys e Sbjcts, exibindo tanto as sequências corretas quanto as que possuem divergências.

```
>NR_169245.1 Homo sapiens colorectal cancer associated 1 (COLCA1), transcript
variant 6, long non-coding RNA
Length=5372

 Score = 9621 bits (10669),  Expect = 0.0
 Identities = 5372/5407 (99%), Gaps = 35/5407 (1%)
 Strand=Plus/Plus

Query  1     GAGTCCCTTCCTTCTCCGCCTGGCCAGGTGTGGCTTCTGGGCAGGCTCCGACCTCTGCGT  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     GAGTCCCTTCCTTCTCCGCCTGGCCAGGTGTGGCTTCTGGGCAGGCTCCGACCTCTGCGT  60

Query  61    GCCCTTGGTCTGGAAGCCAGCCCGGGAGCAAGCGGTGAGGTTTGGCCAGCCCCGTCCTGG  120
             ||||||||||||||||||||||||||||||||||||
Sbjct  61    GCCCTTGGTCTGGAAGCCAGCCCGGGAGCAAGCGG----------------------------  95

Query  121   GCCGGCGAGGAAAAGCCGTGGAAACACACCCGGATTGAAATGCCCCTGGCCCGCCTGACC  180
                       ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  96    ----------AAAAGCCGTGGAAACACACCCGGATTGAAATGCCCCTGGCCCGCCTGACC  145
```

Figura 5. Resultado da busca NCBIWWW salvo no arquivo txt [Dos Autores].

As divergências para serem procuradas e salvas corretamente, devem seguir o método desenvolvido com aplicação da lógica de programação, que busca os intervalos corretos das trincas de bases nitrogenadas. A Figura 6 exibe a lista com os intervalos possíveis para as trincas das bases nitrogenadas.

```
lista_trinca_index = [0, 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36, 39, 42, 45, 48, 51, 54, 57] #possíveis trincas
```

Figura 6. Lista com as possíveis trincas com os indexs iniciais [Dos Autores].

Se o início e fim da “Query” da divergência estiver no início = 2 e fim = 4, o sistema tem que procurar na lista de trincas onde está posicionados o início e fim, e estabelecer o início exato que deverá começar a trinca. No exemplo o sistema estabelece que o início deve ser no índice 0 e o fim no índice 5, pois faz a lógica aplicada verificar se o início é maior ou igual ao valor do índice que está sendo executado. O “while” da condição continua até entrar no “else”, estabelecendo que o início deve ser o valor da

variável no índice - 1 que o “while” estiver executando. Essa mesma lógica é aplicada para definir o fim da Query. A Figura 7 exibe a lógica aplicada para obter o início e fim da divergência.

```
        #encontrar o início
        m = 0
        while True:
            if lista_aux_sbjct4[0] >= lista_trinca_index[m]:
                m += 1
            else:
                inicio_divergencia = lista_trinca_index[m - 1] #definindo início
                #encontrar o fim
                n = 0
                while True:
                    if lista_aux_sbjct4[1] >= lista_trinca_index[n]:
                        n += 1
                    else:
                        final_divergencia = lista_trinca_index[n] - 1 #definindo fim
                        break
                break

    i = inicio_divergencia
    f = final_divergencia
```

Figura 7. Lógica de Programação aplicada no Método [Dos Autores].

4.5. Avaliação

Classificado como uma fase crítica do processo de análise de sequência. É necessária a cooperação de especialistas nos dados, responsáveis pelas decisões. As validações do Software, foram realizadas com a conexão do servidor NCBIWWW, na qual conseguiu encontrar o mesmo arquivo que foi inserido pelo usuário com 100% de alinhamento com ele.

4.6. Aplicação

Essa fase exibe os resultados encontrados na análise de sequência realizado no software. A Figura 8 apresenta o resultado encontrado na análise de sequência, quando a Query pesquisada, possui asterisco (*) no intervalo do início ao fim, e suas divergências.

```
Nome Arquivo: Arquivo >NG_027686.2 Query 3096 --> Divergências:

|                    SEQUÊNCIA QUERY                     |                    SEQUÊNCIA SBJCT                     | INÍCIO DA QUERY | FIM DA QUERY
ATGGAATCTTACTCTGTCACCGGG-CTGGAGTGCAGTGGCATGATCTCAGCTCACTGCAA | ATGGAGTCTTGCTTTGTCGCCCAGTCTGGAGTGAAGTGGTGCGATCTCGGCTCACTGCAA |      3096       |     3154

|     TEM ASTERISCO ?     | POSIÇÃO SEQUÊNCIA DAS BASES | TRINCA ENCONTRADA
  ASTERISCO ENCONTRADO    |            3135             | TGA

|     APRESENTANDO A(S) TRINCA(S) E AMINOÁCIDO(S)
['POSSÍVEIS COMBINAÇÕES: ', 'TGC', 'C', 'TGG', 'W', 'TGT', 'C']

                    APRESENTANDO DIVERGÊNCIAS
| INÍCIO QUERY | FIM QUERY  | TRINCA SEQUÊNCIA QUERY | TRINCA SEQUÊNCIA SBJCT

|      5       |     5      |          GAA           |          GAG

|      10      |     10     |          TAC           |          TGC

|      13      |     13     |          TCT           |          TTT

|      18      |     18     |          ACC           |          GCC

|      21      |     22     |          GGG           |          CAG

|      24      |     24     |          -CT           |          TCT

|      33      |     33     |          CAG           |          AAG

|      39      |     41     |          CAT           |          TGC

|      48      |     48     |          AGC           |          GGC
```

Figura 8. Resultado da Análise com Asterisco (*) na Query Analisada [Dos Autores].

A resposta da conexão com servidor NCBIWWW exibido na Figura 4, possui um tempo estimado em torno de 20 minutos. Pode haver alterações no tempo de resposta,

dependendo das características de processamento do servidor e da relação da quantidade de bases nitrogenadas do arquivo submetido.

Ao realizar a transformação de bases nitrogenadas para aminoácidos, encontramos os asteriscos, que são possíveis mutações na trinca das bases. A Figura 9 apresenta os asteriscos encontrados.

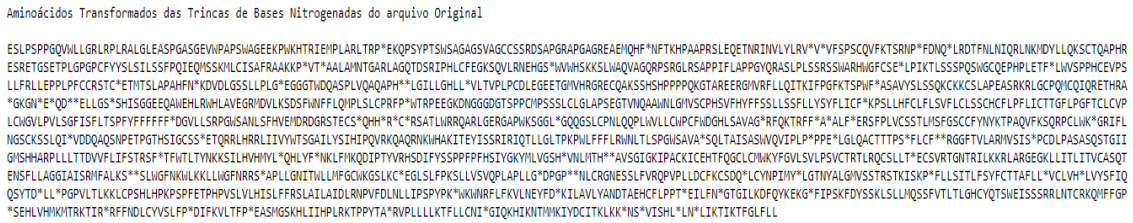


Figura 9. Aminoácidos e os asterisco encontrados [Dos Autores].

5. Conclusões

No transcorrer deste trabalho, ocorreu um empenho para elaborar um sólido fundamento teórico sobre o assunto tratado. Com o conjunto de informações selecionadas foi possível a compreensão pertinente das tecnologias que foram utilizadas no desenvolvimento do Software proposto.

O framework Django contribuiu no desenvolvimento do software, agregando praticidade, agilidade e consistência. Houve a necessidade de seguir as orientações e convenções do Django para o desenvolvimento do sistema.

A biblioteca Biopython possibilitou encontrar as mutações e divergências, utilizando os pacotes disponíveis. Com o módulo NCBIWWW Blast do pacote Bio.Blast conseguiu-se a conexão com o servidor NCBI e a obtenção de respostas, retornando as regiões de similaridade entre sequências biológicas pesquisadas com a base de dados do servidor.

A realização deste projeto apresenta uma integração das áreas de conhecimentos distintos: computação e saúde. É uma ferramenta que pode auxiliar e beneficiar o resultado. A ferramenta desenvolvida, permite maior agilidade na identificação de *proband* com mutações, exibindo as posições que estão ocorrendo os asteriscos e divergências no sequenciamento genético. Aos profissionais de saúde, possibilitará compreensão mais ágil e auxíliará na tomada de decisão com relação as posições e aminoácidos mutados.

Referências

Andrade, G. F., Monteiro, A. F. de S., Menezes, A. Dias, Macedo, M. V. V., Santos, V. P. S. (2020). Mapas Auto Organizáveis e Algoritmos Genéticos Aplicados no Agrupamento e Ranqueamento de Criminosos. Disponível em: http://revistas.poli.br/index.php/repa/article/view/1301/579. Maio/2020.

Araújo, D.; Lima, D.; Campos, P.; Azevedo, V.; Barbosa, J. (2019). Como As Tecnologias De Informação E Comunicação Podem Revolucionar A Saúde E A Medicina. Revista Científica E-Locução, V. 1, N. 15, P. 23, 17 Jul. 2019. Disponível em:http://periodicos.faex.edu.br/index.php/e-Locucao/article/view/187/156. 04/2020.

Biopython (2020a). Biopython Tutorial and Cookbook. Disponível em: http://biopython.org/DIST/docs/tutorial/Tutorial.pdf /. Junho/2020.

Biopython (2020b). Python Tools for Computational Molecular Biology. Disponível em: https://biopython.org/. Julho (2020)

Booch, g., Rumbaugh, j.; Jacobson, I. (2005) Uml: Guia do usuário. Tradução de Fábio Freitas da Silva e Cristina de Amorim Machado. – Rio de Janeiro: Elsevier, 2005 – 9ª Reimpressão.

Caldas, E. (2019). CRISP-DM – Cross Industry Standard Process for Data Mining. Disponível em: https://www.linkedin.com/pulse/crisp-dm-cross-industry-standard-process-data-mining-edmar-caldas. Maio/2020.

Cardoso, N. S., Bispo, T. M. da S. (2019). Um estudo comparativo entre os principais frameworks de desenvolvimento web em linguagem python. Disponível em: http://bdta.ufra.edu.br/jspui/bitstream/123456789/541/1/Um%20Estudo%20Comparativo%20Entre%20os%20Principais%20Frameworks%20de%20Desenvolvimento%20Web%20em%20Linguagem%20Python.pdf. Maio/2020.

Castro, L. R. (2015) Análise de Exômica em pacientes portadores de cardiomiopatia hipertrófica. Disponível em: https://www.teses.usp.br/teses/disponiveis/98/98131/tde-26042016-090654/publico/TeseLaraReinel.pdf

Cordeiro, H. G. (2019) Análise Da Responsividade De Células De Câncer Colorretal Ao Vemurafenibe E Influência Em Proteínas Fosfatases E Vias De Sinalização Do Tgf Beta E Notch1. Disponível em http://repositorio.unicamp.br/bitstream/REPOSIP/335477/1/Cordeiro_HelonGuimaraes_M.pdf. Abril/2020.

Gandra (2020). Inca alerta sobre cuidados para prevenir o câncer colorretal. Disponível em: https://agenciabrasil.ebc.com.br/saude/noticia/2020-03/inca-alerta-sobre-cuidados-para-prevenir-o-cancer-colorretal. Maio/2020.

Guedes, V. R., Bueno, N. F., Oliveira, V. V., Pranchevicius, M. C da S. (2019). Avaliação da expressão imunoistoquimica da proteína p53 no adenocarcinoma colorretal - revisão bibliográfica. Disponível em: https://sistemas.uft.edu.br/periodicos/index.php/patologia/article/view/6830/15015. Maio/2020.

Guedes, T., Ocanã, K., Oliveira, D. (2020). SciPhyloMiner: um Workflow para Mineração de Dados Filogenômicos de Protozários. Disponível em: https://sol.sbc.org.br/index.php/bresci/article/view/9924/9810. Junho/2020.

Inca (2019). Estimativa | 2020 Incidência de Câncer no Brasil. Disponível em: https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//estimativa-2020-incidencia-de-cancer-no-brasil.pdf. Maio/2020.

Inca (2020). Como se comportam as células cancerosas? Disponível em: https://www.inca.gov.br/como-se-comportam-celulas-cancerosas. Maio/2020.

Junior, V. L. de A. (2014). Utilização de Técnicas de Dados Não estruturados para Desenvolvimento de Modelos Aplicados ao Ciclo de Crédito. Disponível em: https://tede2.pucsp.br/bitstream/handle/18150/1/Valter%20Lacerda%20de%20Andrade%20Junior.pdf. Maio/2020.

Lappalainen, T., Scott, A. J., Brandt, M., Hall, I. M. (2019) Genomic analysis in the age of human genome sequencing. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6532068/pdf/nihms-1522722.pdf. Novembro/2020.

Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Bretscher, A., Ploegh, H., and Amonm, A. (2014). Princípios de bioquímica de Lehninger. Artmed Editora, 7 edition.

Medicine, U. S. N. L. (2020a). What is a gene? Disponível em https://ghr.nlm.nih.gov/primer/basics/gene/. Abril/2020.

Medicine, U. S. N. L. (2020b). Our Mission. Disponível em: https://www.ncbi.nlm.nih.gov/home/about/mission/. Junho/2020.

Nelson, D. L. and Cox, M. M. (2014). Princípios de bioquímica de Lehninger. Artmed Editora, 6 edition. Abril/2020.

Paula, M. G de (2018). Modelagem e informatização do processo de análise do Sequenciamento de exoma. Disponível em: https://www.teses.usp.br/teses/disponiveis/59/59143/tde-22112018-173618/publico/MarceloGomes_Revisada.pdf. Dezembro/2020.

Python, S. F. (2020). What is python? executive summary. Disponível em https://www.python.org/doc/essays/blurb/. Maio/2020.

Santos, V. S. dos (2020). "Ácidos Nucleicos". Disponível em: https://brasilescola.uol.com.br/biologia/acidos-nucleicos.htm. Abril/2020.

Song, N., Lee, J., Cho, S., Kim, J., Oh, J. H., Shin, A. (2019). Evaluation of gene-environment interactions for colorectal cancer susceptibility loci using case-only and casecontrol designs. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6918639/pdf/12885_2019_Article_6456.pdf. Maio/2020.

Tenesa, Albert Et Al., (2008). Genome-wide association scan identifies a colorectal câncer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2778004/. Maio/2020.

Teng, H., Gao, R., Qin, N., Jiang, X., Ren, M., Wang, Y., Wu, S., Zhao, J., Qin, H., (2018). Identification of recurrent and novel mutations by whole-genome sequencing of colorectal tumors from the Han population in Shanghai, eastern China. Disponível em: https://www.spandidos-publications.com/10.3892/mmr.2018.9563. Maio/2020.

Zhang, X., Wu, K., Huang, Y., Xu, L., Li, X., Zhang, N. (2020). Promoter Hypermethylation of CHODL Contributes to Carcinogenesis and Indicates Poor Survival in Patients with Early-stage Colorectal Cancer. Disponível em: https://www.jcancer.org/v11p2874.pdf. Abril/2020.