PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO - PRPPG
MESTRADO EM MEIO AMBIENTE E
DESENVOLVIMENTO REGIONAL - MMADRE

# UTILIZAÇÃO DE METODOLOGIAS PARA ANÁLISE COMPARATIVA DE SEQUÊNCIAS NUCLEOTÍDICAS DE UM GENE RELACIONADO AO CÂNCER DE PELE

**ANTONIO ALVES DOS SANTOS NETO**

Presidente Prudente – SP
2015

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
MESTRADO EM MEIO AMBIENTE E
DESENVOLVIMENTO REGIONAL

# UTILIZAÇÃO DE METODOLOGIAS PARA ANÁLISE COMPARATIVA DE SEQUÊNCIAS NUCLEOTÍDICAS DE UM GENE RELACIONADO AO CÂNCER DE PELE

**ANTONIO ALVES DOS SANTOS NETO**

Dissertação apresentada a Pró-Reitoria de Pesquisa e Pós-Graduação, Universidade do Oeste Paulista, como parte dos requisitos para obtenção do título de Mestre em Meio Ambiente e Desenvolvimento Regional.

Orientador: Prof. Dr. Antonio Fluminhan Jr.

Presidente Prudente – SP
2015

570.285 Santos Neto, Antonio Alves dos.
S237i Utilização de metodologias para análise comparativa de sequências nucleotídicas de um gene relacionado ao câncer de pele / Antonio Alves dos Santos Neto. – Presidente Prudente, 2015.
90 f.: il.

Dissertação (Mestrado em Meio Ambiente e Desenvolvimento Regional) - Universidade do Oeste Paulista – Unoeste, Presidente Prudente, SP, 2015.
Bibliografia.
Orientador: Antonio Fluminhan Junior.

1. Bioinformática. 2. Mutações. 3. Câncer de pele. 4. Gene p53. I. Título.

**ANTONIO ALVES DOS SANTOS NETO**

# UTILIZAÇÃO DE METODOLOGIAS PARA ANÁLISE COMPARATIVA DE SEQUÊNCIAS NUCLEOTÍDICAS DE UM GENE RELACIONADO AO CÂNCER DE PELE

Dissertação apresentada a Pró-Reitoria de Pesquisa e Pós-Graduação, Universidade do Oeste Paulista, como parte dos requisitos para obtenção do título de Mestre em Meio Ambiente e Desenvolvimento Regional.

Presidente Prudente, 08 de abril de 2015.

**BANCA EXAMINADORA**

_____
Prof. Dr. Antonio Fluminhan Júnior
Universidade do Oeste Paulista – UNOESTE
Presidente Prudente - SP

_____
Prof. Dr. Milton Hirokazu Shimabukuro
Universidade Estadual Paulista “Júlio de Mesquita Filho” - UNESP
Presidente Prudente - SP

_____
Prof. Dr. Hamilton Mitsugu Ishiki
Universidade do Oeste Paulista – UNOESTE
Presidente Prudente - SP

## DEDICATÓRIA

Dedico este Mestrado a minha esposa Erika e ao nosso filho Enthony, pelo incentivo e apoio em todas as minhas escolhas e decisões.

As vitórias desta conquista dedicam com todo meu amor, unicamente, a vocês!

Parabéns!

## AGRADECIMENTOS

Agradeço a Deus, pelas oportunidades e desafios planejados para mim. Tu és causa primária de todas as coisas e nada se realiza sem a tua vontade.

Ao meu orientador Dr. Antonio Fluminhan Junior agradeço profundamente pelo tempo dedicado, os debates sobre as estratégias, o planejamento e as contínuas revisões e reuniões para que todo esse trabalho ficasse pronto. Sem sua vasta experiência e excelente orientação eu nunca teria chegado até aqui.

A Drª Ana Cristina de Oliveira Lima por viabilizar financeiramente a realização desta conquista.

Ao meu amigo e colega de trabalho Dr. Flávio Alberto Oliva por me incentivar a encarar o desafio de cursar o mestrado.

A minha querida esposa Erika pela paciência e compreensão em conviver comigo durante estes dois anos de muito trabalho e momentos estressantes.

A minha sogra que teve que delongar sua jornada cuidando do meu filho por mais tempo durante o período em que eu estava nas aulas.

Aos meus pais que me ensinaram desde criança que é preciso muito esforço para alcançar os objetivos com dignidade.

A todos os meus parente e amigos que acreditaram e torceram por esta conquista e também por aqueles que direta ou indiretamente colaboraram com o meu crescimento.

“Elevo os meus olhos para os montes; de onde me vem o socorro?
O meu socorro vem do Senhor, que fez os céus e a terra”.
Salmos 121:1-2.

## RESUMO

### UTILIZAÇÃO DE METODOLOGIAS PARA ANÁLISE COMPARATIVA DE SEQUÊNCIAS NUCLEOTÍDICAS DE UM GENE RELACIONADO AO CÂNCER DE PELE

O gene p53 é considerado um dos genes mais intensamente estudados do ponto de vista genômico e está relacionado com os tipos mais comuns de cânceres de pele, os quais mostram uma tendência de aumento da incidência em populações humanas modernas. Sequencias de DNA deste gene estão disponíveis em bancos de dados genômicos com acesso público na internet e fornecem informações úteis para uma grande variedade de aplicações. Esta pesquisa teve como objetivo empregar métodos de análise comparativa das sequências de nucleotídeos para a avaliação de mutações no gene p53 em diferentes países, e avaliar a aplicabilidade das ferramentas de bioinformática disponíveis no ambiente web. As investigações foram realizadas a fim de verificar as mutações mais frequentes no gene p53 já depositadas em bancos de dados genômicos humanos localizados nos Estados Unidos (NCBI), Europa (EBI) e Japão (DDBJ). Na base de dados NCBI foram encontradas 194 sequencias mutantes para o referido gene, sendo que seis destas sequências eram descritas como provenientes de pessoas com câncer de pele. Na base de dados EBI foram encontradas 62 sequencias mutantes do gene p53 em humanos, sendo que 50 eram provenientes de pacientes com câncer de pele. Na base de dados DDBJ foram localizadas 4075 sequencias mutantes para o mesmo gene, mas não identificadas como provenientes de pacientes com câncer. A seqüência original do gene p53 foi comparada com as sequências mutantes oriundas de pacientes com diversos tipos de cânceres de pele, por meio das ferramentas de bioinformática CLUSTALW e MAFFT. Estudos comparando as sequências mutantes permitiram estabelecer relações entre as mutações mais freqüentes e que podem estar relacionados com a patologia nos diferentes países. Os resultados mostraram a viabilidade da abordagem proposta, e permitiram a identificação de características comuns entre os mutantes analisados. Foi possível identificar as regiões gênicas mais sensíveis às alterações nas sequências de nucleotídeos no gene p53, e que podem estar relacionados com a origem da patologia. Também foi possível identificar que as mutações comumente ocorrem em regiões bem próximas, tanto para as sequencias oriundas do banco de dados NCBI quanto para aquelas coletadas do banco de dados EBI.



Palavras-chave: Bioinformática, mutações, câncer de pele, gene p53.

## ABSTRACT

### UTILIZATION OF METHODOLOGIES FOR COMPARATIVE ANALYSIS OF NUCLEOTIDE SEQUENCES OF A GENE RELATED TO THE SKIN CANCER

The p53 gene is considered one of the most intensively studied genes from the genomic point of view and is related to the most common types of skin cancers, which show a trend of increasing incidence in modern human populations. DNA sequences of this gene are available in genomic databases with public access on the Internet and provide useful information for a wide range of applications. This research aimed to employ methods of comparative analysis of nucleotide sequences for evaluating mutations in the p53 gene in different countries, and evaluate the applicability of bioinformatics tools available on the web environment. Investigations were carried out to verify the most frequent mutations in the p53 gene whose sequences are deposited in human genomic databases located in the United States (NCBI), Europe (EBI) and Japan (DDBJ). In the NCBI database we found 194 mutant sequences of the p53 gene, and six of them were described as coming from people with skin cancer. In EBI database 62 mutant sequences of the p53 gene were found, and 50 were from patients with skin cancer. In DDBJ database we found 4075 mutant sequences for the same gene, but not identified as coming from patients with cancer. The original sequence of the p53 gene was compared with the mutant sequences derived from patients with various types of skin cancers, through the CLUSTALW and MAFFT bioinformatics tools. Studies comparing the mutant sequences allowed to establish relationships among the most frequent mutations and that might be related to the disease in different countries. The results showed the feasibility of the proposed approach, and allowed the identification of common characteristics between the analyzed mutants. It was possible to identify the gene regions most sensitive to changes in nucleotide sequences in the p53 gene, which might be related to the origin of the pathology. It was also possible to identify that the mutations commonly occur in very close areas, both for the sequences of the NCBI database as to those collected from EBI database.



Keywords: Bioinformatics, mutations, skin cancer, p53 gene

## LISTA DE SIGLAS

AIA – Avaliação de Impacto Ambiental

BLAST – Basic Local Alignment Search Tool

CFCs - Clorofluorcarbono

$CO_2$ – Dióxido de Carbono

DDBJ – DNA Data Bank of Japan

DNA – Ácido Desoxirribonucleico

EBI – European Bioinformatics Institute

EMBL – European Molecular Biology Laboratory

EUA – Estados Unidos da América

GSDB – Gene Sequence Database

INCA – Instituto Nacional de Câncer

Km – Quilômetros

LCGBI – Laboratório de Citogenômica e Bioinformática

Mbps - mega bits por segundo

NCBI – National Center for Biotechnology Information

$O_2$ – Oxigênio

$O_3$ – Ozônio

PCR – Reação em Cadeia de Polimerase

UNOESTE – Universidade do Oeste Paulista

UV – Ultravioleta

Web – Sistema de hipertextos que opera através da internet

## LISTA DE FIGURAS

## LISTA DE TABELAS

**SUMÁRIO**

# 1 INTRODUÇÃO

Segundo Santos (2007), os dramas ambientais encontrados recentemente na natureza foram causados por gerações e gerações que desconheciam o equilíbrio homem/ambiente e construíram um modelo de desenvolvimento predatório. O impacto da espécie humana sobre o meio ambiente tem sido comparado, por alguns cientistas, às grandes catástrofes do passado geológico da terra.

Em função do aumento da população, surgiu a necessidade de se produzir cada vez mais, submetendo o meio ambiente a uma agressão que está provocando o declínio cada vez mais acelerado de sua qualidade e de sua capacidade para sustentar a vida. O uso de combustíveis fósseis tem aumentado a concentração de dióxido de carbono ($CO_2$) na atmosfera, dando lugar a um aumento da temperatura global na terra (SANTOS, 2007).

De acordo com Dessler (2000) o homem moderno criou substâncias artificiais que tem destruído a camada de ozônio, aumentando a penetração da radiação ultravioleta do tipo B, altamente prejudicial para os seres vivos. A camada de ozônio é uma região da atmosfera terrestre localizada em torno de 20 a 30 km de altura e tem importância fundamental para a vida no planeta terra. Ainda segundo o autor, é ela que absorve a radiação ultravioleta do tipo B emitida pelo sol e não permite que esta radiação, prejudicial à vida, chegue até a superfície terrestre (DESSLER, 2000).

Segundo Dessler (2000) a principal consequência da destruição da camada de ozônio tem sido o aumento da incidência do câncer de pele, já que os raios ultravioletas são mutagênicos. Diante desta realidade, tem surgido a necessidade de se criar metodologias para a avaliação das consequências dos impactos ambientais decorrentes de atividades humanas ou provocados por fenômenos naturais, em especial ao nível biológico.

Segundo Souza (2004) a determinação da estrutura do ácido desoxirribonucleico (DNA) por James Watson e Francis Crick em 1953, fez surgir uma nova extensão da ciência conhecida como biologia. Ainda segundo o autor, o computador contribuiu significativamente no sequenciamento do genoma humano, criando um marco importante na história da evolução do conhecimento humano.

De acordo com Geraque (2003) a rede mundial de computadores, mais

conhecida como internet, tem contribuído ativamente com pesquisadores do mundo inteiro proporcionando-lhes acesso gratuito a banco de dados com informações sobre bilhões de genes e possibilitando a comparação entre eles. Ainda segundo o autor, essa evolução proporcionou à bioinformática ferramentas úteis para a realização de pesquisas nesta área do conhecimento. O autor também afirma que os dados biológicos armazenados em banco de dados com acesso público, tais como sequências de DNA e de proteínas tem viabilizado também o trabalho de muitos cientistas, principalmente nos estudos de comparações de sequências de nucleotídeos para diversas finalidades, tais como: identificação de mutações ou polimorfismos genéticos, estudos evolutivos e filogenéticos. A disponibilidade dessas informações e a facilidade de acesso à internet revolucionou a forma como as pesquisas, desta área, são realizadas na atualidade, reduzindo significativamente o tempo destinado aos trabalhos (GERAQUE, 2003).

De acordo com Martinez et al. (2006) algumas das alterações na estrutura do DNA têm sido atribuídas à exposição excessiva aos raios ultravioletas (UV), em especial ao ultravioleta tipo B (UVB), e como consequência direta o aumento da incidência de canceres cutâneos. Ainda segundo o autor, quando ocorrem danos no DNA, desencadeiam-se nas células mecanismos bioquímicos capazes de reparar tais lesões. Para que este processo ocorra é necessário que a célula pare o ciclo celular a fim de não perpetuar a mutação, essa função conhecida como: apoptose (morte celular programada) está atribuída à proteína TP53, normalmente referida como a “guardiã do genoma”. (MARTINEZ et al., 2006)

## 1.1 Objetivos

### 1.1.1 Objetivos gerais

Os objetivos gerais deste trabalho envolvem a utilização de metodologias no Laboratório de Citogenômica e Bioinformática (LCGBI) da UNOESTE visando à análise comparativa de sequências de nucleotídeos em genes marcadores de impactos ambientais. Para tanto, serão realizadas pesquisas com o objetivo de averiguar as mais frequentes mutações do gene p53 nos bancos de dados do Instituto Europeu de Bioinformática (EBI), Centro Nacional de Informações sobre Biotecnologia (NCBI) e no Banco de Dados de DNA do Japão (DDBJ)

respectivamente localizados na Europa, Estados Unidos e Japão que são os mais importantes bancos de dados genômicos de acesso público.

### 1.1.2 Objetivos específicos

- Viabilizar a realização de pesquisas com foco no estudo de impactos ambientais, de modo a disponibilizar técnicas mais avançadas para avaliação do grau de mutagenicidade e o risco biológico apresentado por uma determinada ação antrópica ou natural.

- Acessar sequências mutantes do gene p53 de pacientes diagnosticados com câncer de pele disponíveis em bancos de dados genômicos de acesso público na internet e compará-las utilizando ferramentas de Bioinformática.

- Avaliar as mutações gênicas mais comuns em populações humanas oriundas de países com diferenças contrastantes em relação aos fatores: grupo étnico, localização geográfica em respeito à latitude e valores de IDH, entre outros.

- Otimizar recursos disponíveis na internet, tais como sequencias mutantes do gene p53 e softwares de análise de sequencias genômicas, comparando-as em uma estação de trabalho no laboratório de Citogenômica e Bioinformática da Unoeste.

- Possibilitar a cooperação entre grupos de pesquisa das áreas de Ciência da Computação, Engenharia e Estatística com as áreas de Biologia e Química, no intuito de criar grupos de pesquisas interdisciplinares de Bioinformática na UNOESTE, buscando estabelecer vínculos interinstitucionais nesta área de pesquisa.

## 1.2 Justificativas

Segundo Karp (2005) o uso de bases de dados de estudo de similaridade de sequências de DNA é uma realidade presente em várias Instituições

que desenvolvem pesquisas em Biologia Molecular. Sem a utilização de programas de busca e análise de sequências seria impossível fazer predições que minimizam o árduo caminho da pesquisa em laboratório.

A colaboração entre diferentes grupos de pesquisa em vários países tem impulsionado a Bioinformática, tecnologia que veio propiciar aos pesquisadores novas habilidades e competências em gerenciar suas informações biológicas. As instituições de pesquisa em Bioinformática têm expandido significativamente sua capacidade. Seus recursos constituem-se em instrumentos estratégicos para a utilização racional dos dados genômicos. Constata-se a existência de grupos de pesquisa em áreas que têm qualificações que se complementam para o desenvolvimento de pesquisas de alto nível no campo da Bioinformática, notadamente nas áreas aplicadas à genômica e à modelagem molecular. Pesquisas de novos genes em diversas áreas têm utilizado e requerido à aplicação de Bioinformática no detalhamento dos produtos gênicos.

A implantação de recursos computacionais com acesso a rede mundial de computadores possibilitou novas formas de experimentação laboratorial. Foi possível desenvolver um aprofundamento de estudos e experimentação que, usando os resultados de experimentos biológicos, possa alimentar análises de sequências de DNA, através das bases de dados acessíveis via Internet.

Na UNOESTE, alguns grupos já produzem, por iniciativas isoladas, trabalhos de iniciação científica e monografias relacionadas com a Bioinformática. Na Faculdade de Informática de Presidente Prudente – FIPP, já foram desenvolvidos trabalhos na área de bioinformática. Atualmente não há trabalhos em andamento nos cursos da FIPP voltados para a Bioinformática. Uma avaliação realista das potencialidades da Universidade em muito poderá contribuir para o caráter multidisciplinar destas atividades, aumentando a relevância e a produtividade nas análises Bioinformáticas. Neste contexto, a UNOESTE vem se preparando para uma inserção significativa na chamada era pós-genômica, de transformação da informação genômica em conhecimento científico e tecnológico.

As ferramentas de análise de sequencias estão disponíveis em vários servidores Web, conferindo uma vantagem para as análises de sequencias disponibilizadas em banco de dados genômicos. O domínio sistemático deste processo pode trazer uma forte contribuição para o incremento dos resultados das pesquisas, as quais podem ser realizadas em diversos locais do mundo,

simultaneamente, fato este que também justifica a realização do presente trabalho.

# 2 REVISÃO DA LITERATURA

## 2.1 Impactos ambientais decorrentes da ação humana

Segundo Santos (2007), o homem construiu um modelo de desenvolvimento predatório que se agravou exponencialmente durante a revolução industrial, e que a partir de então começou a transformar o que há de mais importante para a nossa subsistência na terra como a natureza de sua atmosfera e a qualidade de sua água. A humanidade deve reconhecer que as atividades que impactam o meio ambiente põem em risco a sobrevivência de sua própria espécie.

Ainda de acordo com Santos (2007), o rápido crescimento da população, gerou uma demanda por bens materiais e serviços sem precedentes, que o desenvolvimento tecnológico pretende satisfazer, como a produção de veículos, alimentos e produtos tecnológicos em grande escala, submetendo o meio ambiente a uma agressão que está provocando o declínio cada vez mais acelerado de sua qualidade e de sua capacidade para sustentar a vida.

O autor também afirma, que um dos impactos que o uso de combustíveis fósseis tem produzido sobre o meio ambiente terrestre é o aumento da concentração de dióxido de carbono ($CO_2$) e outros gases responsáveis pelo efeito estufa na atmosfera, dando lugar a um aumento da temperatura média global da Terra. Outros males importantes causados pelo ser humano ao meio ambiente são o uso de agrotóxicos que contaminam regiões agrícolas e interferem no metabolismo do cálcio das aves; a erosão do solo; o crescente problema mundial do abastecimento de água, como consequência do esgotamento dos aquíferos subterrâneos, assim como pela queda na qualidade e disponibilidade da água e a destruição da camada de ozônio (SANTOS, 2007).

De acordo com Christopherson (2012), em função do crescimento populacional, da industrialização acelerada e a escassez de alimentos o homem moderno criou substancias artificiais que tem contribuído para a diminuição da camada de ozônio, aumentando a penetração da radiação altamente prejudicial para os seres vivos. Ainda segundo o autor, a molécula de oxigênio fundamental para a sobrevivência de grande parte da vida na terra é composta por dois átomos de oxigênio e é representado pelo símbolo $O_2$. O ozônio é composto de três átomos de oxigênio e a sua representação é através do símbolo $O_3$ (CHRISTOPHERSON,

2012).

Segundo Christopherson (2012) dentre as substancias prejudicias podemos destacar os clorofluorcarbonetos (ou CFCs), que são moléculas sintéticas de cloro, flúor e carbono e que são estáveis e inertes sob as condições da superfície terrestre. O autor afirma que esta substancia foi muito utilizada em sprays de aerossóis, indústrias eletrônicas e agentes espumantes. Porém as moléculas estáveis de CFCs migram lentamente para a estratosfera, onde a intensa radiação ultravioleta divide-as, liberando os átomos de cloro. Ainda segundo o autor, esse processo produz reações que quebram as moléculas de ozônio e deixam moléculas de gás oxigênio em seu lugar. O efeito é grave, pois um único átomo de cloro decompõe mais de 100.000 moléculas de ozônio. O autor afirma que, desde que foram criados, mais de 22 milhões de toneladas de CFCs foram comercializadas em todo o mundo e, posteriormente, liberadas na atmosfera (CHRISTOPHERSON, 2012).

Segundo Santos (2007), há bastante ozônio nas camadas altas de nossa atmosfera e ele possui tamanho e formato ideal para absorver a energia do sol, na faixa da radiação ultravioleta, prejudicial à vida. O ozônio forma uma camada que impede que grande parte dos raios ultravioletas atinja a terra. Dessler (2000) afirma que vários produtos químicos produzidos pelo homem são transportados até a estratosfera, onde está localizada a camada de ozônio. Conforme o autor, quando a luz do sol atinge algumas dessas moléculas, elas se quebram e produzem outros átomos que atingem as moléculas de ozônio desencadeando uma reação química que transformam o ozônio em oxigênio molecular. O oxigênio não é capaz de absorver a radiação solar que é perigosa para a vida na Terra.

De acordo com Dessler (2000), o ozônio é um gás composto de três átomos de oxigênio atômico, unidos por ligações simples e duplas. Em temperatura ambiente tem coloração azul-pálida, devida à intensa absorção de luz vermelha, atingindo coloração azul-escura quando transita para o estado líquido, situação em que adquire propriedades explosivas.

Segundo Dessler (2000) o ozônio é produzido naturalmente na estratosfera pela ação fotoquímica dos raios ultravioletas sobre as moléculas de oxigênio. Ainda segundo o autor, a concentração de ozônio é resultado de um equilíbrio de produção e destruição gerando camadas de alta e baixa concentração que atingem números máximos numa faixa de 30 Km, chamada de camada de

ozônio. O autor afirma que esta faixa de ozônio tem cerca de 15 km de espessura e sua constituição, há cerca de 400 milhões de anos, permitiu o desenvolvimento de vida na terra, já que impede a passagem dos raios UVB (DESSLER, 2000).

Segundo Rocha (2009) esta camada tem capacidade de impedir que a radiação ultravioleta do tipo B, prejudicial à vida, atinja a superfície da terra. Ainda segundo o autor, a radiação solar é a energia que emana do sol e está distribuída em diversas ondas: desde o infravermelho até o ultravioleta (UV). O autor também afirma que os raios UV são divididos em UV-C totalmente absorvido na atmosfera terrestre; o UV-A que não é absorvido pela atmosfera e o UV-B que é absorvido pela camada de ozônio (ROCHA, 2009).

De acordo com Dessler (2000), uma das consequências da destruição da camada de ozônio tem sido o grande aumento da incidência do câncer de pele, já que os raios UV são mutagênicos. Segundo o autor, a exposição excessiva à radiação ultravioleta (UV), em especial ao ultravioleta tipo B (UVB), tem sido associada ao aumento do risco para o desenvolvimento dos canceres cutâneos, pois pode causar mutações genicas no ácido desoxirribonucleico (DNA) dos queratinócitos. A falha no reparo dessas alterações genicas pode levar a um crescimento celular desordenado e formação de tumor. Além disso, o autor menciona que a radiação UV tem grande efeito sobre o sistema imune cutâneo, induzindo a um estado de imunossupressão local que impede a rejeição do tumor neoformado (DESSLER, 2000).

Segundo Melo (2012), várias empresas já incluíram a questão do meio ambiente na sua agenda, entretanto a maioria dos problemas ambientais elementares ainda persiste em função do seu tratamento requerer uma mudança nos meios de produção industrial e de consumo, bem como de nossa organização social e de nossas vidas pessoais.

Segundo o Instituto Nacional Do Câncer (Brasil) (2014) o câncer de pele não melanoma é o tipo mais frequente no Brasil e corresponde a 25% de todos os tumores malignos registrados no país. Apresenta altos percentuais de cura, se for detectado precocemente. Entre os tumores de pele, o tipo não melanoma é o de maior incidência e mais baixa mortalidade. O Instituto afirma também que o câncer de pele é mais comum em pessoas com mais de 40 anos, sendo relativamente raro em crianças e negros, com exceção daqueles já portadores de doenças cutâneas anteriores. Pessoas de pele clara, sensível à ação dos raios solares, ou com

doenças cutâneas prévias são as principais vítimas.

Ainda segundo o (INCA), como a pele - maior órgão do corpo humano - é heterogênea, o câncer de pele não melanoma pode apresentar tumores de diferentes linhagens. Os mais frequentes são carcinoma basocelular, responsável por 70% dos diagnósticos, e o carcinoma epidermóide, representando 25% dos casos. O carcinoma basocelular, apesar de mais incidente, é também o menos agressivo.

Para o ano de 2012 o INCA projetou cerca de 134 mil novos casos somente de canceres cutâneos, sendo aproximadamente 62 mil homens e 71mil mulheres. Em 2013, no dia nacional de combate ao câncer, foi divulgada a projeção para o ano de 2014 de 182 mil novos casos somente do tipo não melanoma. Um aumento de mais de 35% em apenas dois anos.

O instituto tem lançado várias campanhas em atendimento à Política Nacional de Prevenção e Controle do Câncer do Ministério da Saúde com o objetivo de conscientizar as pessoas sobre o risco da exposição solar excessiva.

Diante desta realidade tem surgido a necessidade de se criar metodologias para a avaliação dos impactos ambientais, decorrentes de atividades humanas ou provocados por fenômenos naturais, em especial ao nível biológico que pode ter como consequência a incidência de doenças ou até mesmo o risco de morte.

## 2.2 Avaliação de impacto ambiental

Segundo Sanchez (1995), a Avaliação de Impacto Ambiental (AIA) deve ser compreendida como instrumento de planejamento, isto é, como uma atividade técnico-científica que tem por finalidade identificar, prever e interpretar os efeitos de uma determinada ação humana sobre o ambiente.

Segundo Carvalho (2010), a Avaliação de Impacto Ambiental foi criada em 1969 pela lei da política ambiental dos Estados Unidos e tornou-se modelo para as legislações similares em diversos países. No fim da década de 70, princípio da sua utilização, a (AIA) analisava apenas os meios físico e biótico, já na década de 80 passou a avaliar os aspectos sociais e de saúde, análise de risco e incluir a participação pública. Hoje é possível afirmar que houve um desenvolvimento significativo das metodologias de AIA.

Esses sistemas de AIA variam muito entre os vários países. Alguns são leis, normas ou estatutos, que são exigidos pelas autoridades antes da permissão de implementação de um projeto. Em outros casos, apenas diretrizes sobre AIA foram estabelecidas, impondo algumas obrigações para os órgãos governamentais.

De acordo com a CETESB (2014), no Brasil, já na década de 70, projetos de grande porte, financiados por organismos multilaterais, foram submetidos à Avaliação de Impacto Ambiental, como por exemplo, a Usina Hidrelétrica de Sobradinho, a Usina Hidrelétrica de Tucuruí, etc. Tais experiências promoveram a inclusão da AIA como um dos instrumentos da Política Nacional de Meio Ambiente, Lei nº 6938/81, em associação ao licenciamento das atividades utilizadoras dos recursos ambientais, consideradas efetiva ou potencialmente poluidoras. Em 1986, foi editada a Resolução Conama 01/86, estabelecendo as definições, responsabilidades, critérios básicos e as diretrizes para o uso e implementação da avaliação de impacto ambiental, aplicado ao licenciamento ambiental de determinadas atividades modificadoras do meio ambiente (CETESB, 2014).

Segundo Souza (2004), o avanço tecnológico proporcionou o surgimento de novas ferramentas e métodos de avaliação. De acordo com o autor, desde que James Watson e Francis Crick, em 1953, determinaram a estrutura do Ácido Desoxirribonucleico (DNA), surgiu uma nova extensão da ciência conhecida como Biologia Molecular que não parou de evoluir. O autor também afirma que uma grande quantidade de sequencias de DNA passou a ser sintetizadas e analisadas em escala crescente, graças à aliança com a informática. Com capacidade de processamento cada vez maior, o computador proporcionou até mesmo o sequenciamento completo do genoma humano, iniciado em 1985 e finalizado em 2000 (PGH – Projeto Genoma Humano) um marco importante na evolução do conhecimento científico (SOUZA, 2004).

## 2.3 A contribuição da bioinformática para os estudos genéticos

Segundo Sabbatini (1993), o uso da informática nos campos de estudo da vida é essencial para a interação entre as ciências, propiciando assim um melhor desenvolvimento humano e tecnológico. Conforme o autor, nos últimos anos, a Biologia vem se apropriando com avidez das ferramentas de software e hardware

proporcionadas pela informática. Assim, o computador tornou-se peça chave nas pesquisas que vem sendo realizadas em áreas especializadas como a Biologia Molecular e o armazenamento de dados biológicos em bancos de dados públicos tem se tornado cada vez mais comum. O autor menciona que a “febre” genômica direcionada para a identificação e o mapeamento dos milhares de genes nas células de inúmeros organismos vivos, que se alastrou pelo mundo, em especial o projeto Genoma Humano, fez com que a Informática Biológica se desenvolvesse rapidamente (SABBATINI, 1993).

A rede mundial de computadores tem contribuído ativamente com pesquisadores do mundo inteiro proporcionando-lhes acesso gratuito a banco de dados com informações sobre bilhões de genes e possibilitando a comparação entre eles. Essa evolução proporcionou à bioinformática ferramentas úteis para a realização de pesquisas nesta área do conhecimento.

De acordo com Baxevanis (1998), dados biológicos armazenados em banco de dados com acesso público, tais como sequências de DNA e de proteínas, tem facilitado o trabalho de muitos cientistas e pesquisadores, principalmente nos estudos de comparações de sequências de nucleotídeos. O autor alega que estes estudos são uteis para diversas finalidades, tais como: identificação de mutações ou polimorfismos genéticos, estudos evolutivos e filogenéticos. A disponibilidade dessas informações e a facilidade de acesso através da internet revolucionou a forma como as pesquisas da área de bioinformática são realizadas na atualidade reduzindo significativamente o tempo destinado aos trabalhos de pesquisa relacionados a sequenciamento genético (BAXEVANIS, 1998).

As situações acima descritas indicam claramente uma confluência entre a Biologia, a Ciência da Computação e a Estatística, configurando novas linhas de pesquisa, que foram denominadas: a “Biologia Computacional”, que estuda o desenvolvimento da Tecnologia da Informação para resolver os problemas biológicos, e a “Bioinformática”, que é a aplicação dessa tecnologia no gerenciamento dos bancos de dados biológicos (SETUBAL; MEIDANIS, 1997).

Segundo Gibas e Jambeck (2001), a comparação de sequências de DNA é uma das bases da Bioinformática. De acordo com o autor, os programas de comparações sequenciais mudaram a Biologia Molecular, e a Web passou a possibilitar que bancos de dados públicos, de sequências de genoma, ofereçam serviços por meio de uma interface uniforme para uma comunidade mundial de

pesquisadores. Com um *software* específico, instalado em um computador comum, um biólogo molecular pode, então, comparar uma sequência de DNA desconhecida com a coleção completa de sequências de DNA públicas. Um exemplo desses programas é o BLAST (*Basic Local Alignment Search Tool*) do National Center for Biotechnology Information – NCBI, localizado nos Estados Unidos (GIBAS; JAMBECK, 2001).

De acordo com Zhang e Massen (1997), a capacidade de executar comparações automatizadas de sequências simplifica a atribuição de função para uma nova sequência.

Gibas e Jambeck (2001), afirmam que com o acúmulo de dados de sequências biológicas e análise atenta, conclui-se que a natureza pode ser considerada conservadora. Ainda segundo o autor, uma nova via biossintética pode não ser criada para cada nova sequência descoberta, porém, uma nova funcionalidade pode ser descrita a partir da descoberta de novos genes. Com isso, a comparação de sequência de espécies diferentes possibilita encontrar sequências similares com razoável fidelidade, mas nem sempre total (GIBAS; JAMBECK, 2001).

Segundo Ticona (2003), o conceito básico da comparação de sequências é simples: duas sequências são combinadas aleatoriamente e as semelhanças da combinação são avaliadas e pontuadas. Em seguida, o mesmo procedimento é realizado para outra sequência, e a combinação pontuada novamente, até que a melhor combinação seja encontrada. Sem dúvida é um processo simples, mas a complexidade está na existência de milhares de combinações possíveis (TICONA, 2003).

## 2.4 Alinhamento de sequências

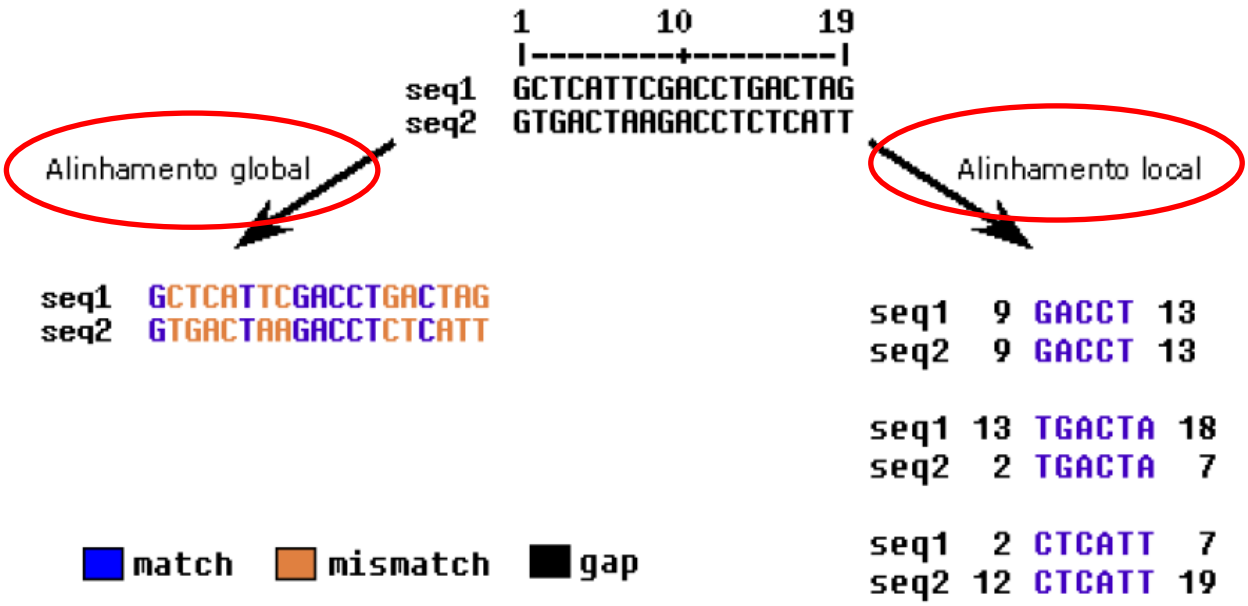
Segundo Prosdocimi (2007), o alinhamento de sequências consiste no processo de comparar duas sequências (de nucleotídeos ou proteínas) de forma a se observar seu nível de identidade. Ainda segundo o autor, essa técnica de comparação de sequências é implementada segundo um conceito de desenvolvimento de programas conhecido como um algoritmo guloso[1] e é um dos pilares de toda a bioinformática. O Autor afirma também, que existem centenas de

[1] Algoritmo guloso é uma técnica de algoritmos para resolver problemas de otimização, sempre realizando a escolha que parece ser a melhor no momento; fazendo uma escolha ótima local, na esperança de que esta escolha leve até a solução ótima global.

aplicações do alinhamento de sequências, tanto na identificação de genes e proteínas desconhecidas, quanto na comparação da ordem de genes em genomas de organismos proximamente relacionados (sintenia), no mapeamento de sequências expressas dentro de um genoma para identificação de genes, na montagem de genomas e em diversas outras aplicações.

De acordo com Prosdocimi et al. (2002), pode-se alinhar duas sequências para descobrir o grau de similaridade entre elas de forma que possa inferir (ou não), alguma propriedade já conhecida da outra. Ainda segundo o autor o alinhamento entre duas sequências pode ser feito de forma global exemplificado na Figura **1**, lado esquerdo, ou de forma local, no lado direito.

Figura 1 – Ilustração do Alinhamento global à esquerda e Alinhamento local à direita




Fonte: Prosdocimi (2007, p.18)

No lado esquerdo da Figura **1** está exemplificado como é realizado um estudo de alinhamento global quando se utiliza dois conjuntos de dados. Duas sequencias são comparadas do início ao final e são atribuídos valores correspondentes à quantidade de caracteres iguais encontrados nas duas sequencias e na mesma posição. As semelhanças estão destacadas com a cor azul e a estas semelhanças são atribuídos valores maiores. A soma de todos os valores aferidos corresponde ao valor de *match* (correspondência) entre as sequencias.

No lado direito da mesma figura está ilustrado como ocorre o

alinhamento local. Neste caso, fragmentos da sequencia 1 são alinhados com fragmentos da sequencia 2. Antes de cada fragmento, tanto da sequencia 1 quanto da sequencia 2, é inserido a posição inicial e final do fragmento em relação a sequencia completa. Também são atribuídos valores para cada correspondência de bases entre ambas sequencias e o valor somatório corresponde ao valor de *match* (correspondência) entre as sequencias.

Segundo Lima (2012), a comparação de sequências biológicas pode ocorrer entre duas sequências (alinhamento em pares) ou entre mais de duas sequências (alinhamento múltiplo). A utilização de duas sequências ou mais pode facilitar a visualização das diferenças e semelhanças entre os conjuntos de sequências como um todo. Uma alternativa ao alinhamento múltiplo seria a execução de várias comparações entre duas sequências. Ao alinhar as sequências em pares, as diferenças e semelhanças globais podem não ser claras. Por outro lado, alinhando-se múltiplas sequências, essas diferenças e semelhanças podem se tornar claras, ou, até mesmo, óbvias. De posse de um alinhamento múltiplo, os biólogos podem então entender, por exemplo, como a evolução atuou, descobrindo a diferença genética entre as espécies (LIMA, 2012).

### 2.4.1 Alinhamento global

O alinhamento global é feito quando compara-se uma sequência de aminoácidos com outra sequencia de aminoácidos ou uma sequencia de nucleotídeos com outra sequencia de nucleotídeos, ao longo de toda a extensão das sequencias.

De acordo com Prosdocimi (2007), o algoritmo Needleman-Wunsch é utilizado para alinhar globalmente pares de sequencias, maximizando o número de alinhamentos individuais (um-para-um) que são iguais (*match*) e minimizando os espaçamentos (*gap*). O autor também afirma que, nesse caso são dados valores em uma matriz de comparação para as similaridades (*matches*), para diferenças (*mismatches*) e para falhas (*gaps*) encontradas durante o alinhamento das sequências. As somas dos valores do alinhamento, de acordo com essa matriz de comparação, resultam num valor, que é um escore de similaridade entre as sequências conforme visto na Figura 2 (PROSDOCIMI, 2007).

### 2.4.2 Alinhamento local

O alinhamento local acontece quando a comparação entre duas sequências não é feita ao longo de toda sua extensão, mas sim através de pequenos sítios destas sequencias. O principal programa utilizado para o alinhamento local de sequências é o BLAST (*Basic Local Alignment Search Tool* ou Ferramenta Básica de Procura por Alinhamento Local), encontrado em http://www.ncbi.nlm.nih.gov/BLAST/. Esse software compreende um conjunto de algoritmos de comparação de sequências montado de forma a explorar toda a informação contida em bases de dados de DNA e proteínas. O programa BLAST foi desenvolvido de modo a aumentar ao máximo a velocidade da busca por similaridade, já que as bases de dados são grandes e vêm crescendo exponencialmente, mesmo correndo o risco de perder um pouco na precisão do resultado (ALTSCHUL; MADDEN; SCHÄFFER, 1997).

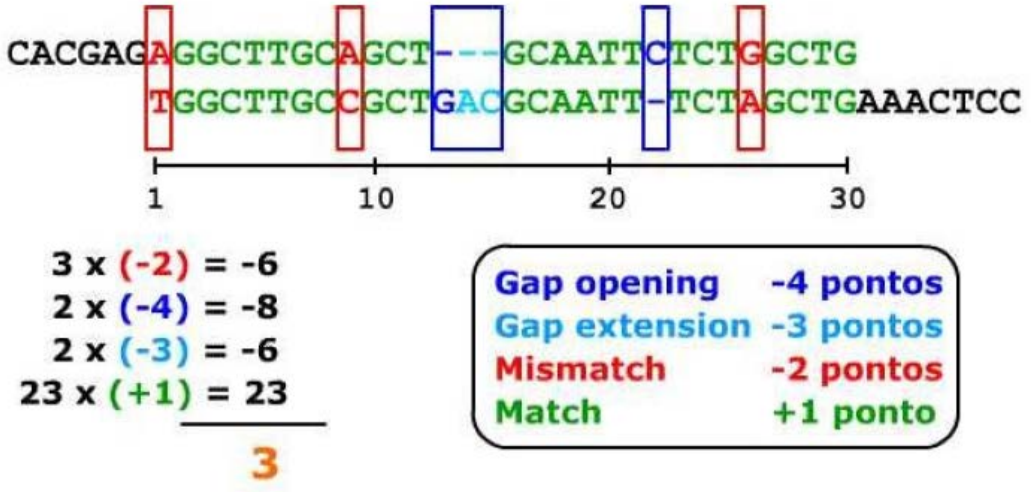
A rapidez da busca deve-se ao fato de que o programa utiliza uma heurística que quebra as sequências de entrada e das bases de dados em fragmentos – as palavras (*words*) – e procura, inicialmente, similaridades entre elas (ALTSCHUL; MADDEN; SCHÄFFER, 1997). Outra vantagem de se utilizar o alinhamento local feito pelo BLAST é que, dessa forma, é possível identificar relações entre sequências que apresentam apenas sítios isolados de similaridade.

O alinhamento de sequências de DNA é feito através da procura de uma região de similaridade entre duas sequências utilizando um algoritmo guloso. Quando essa região é encontrada são dados pontos para similaridades (*match*), diferenças (*mismatches*), abertura de falhas (*gap opening*) e extensão de falhas (*gap extension*) que possam ser encontradas no seu alinhamento. A somatória dos pontos desse alinhamento é chamada de escore do alinhamento. Na Figura 2 o escore do alinhamento é três. Tais escores são contabilizados tanto nos alinhamentos globais quanto nos locais (PROSDOCIMI, 2007).

O score de valor 3 calculado na Figura 2 é a somatório dos pontos encontrados para *gap opening, gap extension, mismatch e match*, conforme visto na imagem. A identificação dos pontos é realizada de acordo com o melhor posicionamento entre as duas sequencias. Gap opening é a quantidade de ‘-‘ inseridos entre as sequencias para que as mesmas ficassem mais bem

posicionadas. Gap extension é a maior quantidade de '-' que foram sequencialmente posicionados. Mismatch é a identificação das trocas dos nucleotídeos. Para o levantamento do pontos são atribuídos valores negativos, exceto para o match que são os nucleotídeos idênticos. Para se chegar ao valor de score, cada ponto é multiplicado pela quantidade de vezes que eles se repetem na extensão de toda a sequencia e na sequencia da multiplicação, os resultados são somados, resultando no valor de score.

Figura 2 – Ilustração de Alinhamento local entre duas sequencias e a somatória de *score*




Fonte: Prosdocimi (2007, p.19)

Para o desenvolvimento de experiências com esse objetivo, têm-se softwares específicos, como é o caso do BLAST, disponível nos sítios da internet dos centros de pesquisa especializados. Assim, o acesso à ferramenta BLAST pode ser feito através das *webpages* do NCBI e do DDBJ (*DNA Data Bank of Japan*). A *webpage* do NCBI disponibiliza também ferramentas de pesquisa do banco de dados HomoloGene, que armazena pares de genes homólogos. O NCBI oferece acesso a uma ampla seleção de ferramentas de análise de genoma baseadas na Web, através da seção *Genomic Biology* (Biologia Genômica). Seu website dispõe também de ampla documentação para auxiliar na aprendizagem do uso de suas ferramentas e banco de dados. Porém, para realizar uma comparação de sequência em banco de dados, a ferramenta BLAST é a mais usada (ALTSCHUL; MADDEN; SCHÄFFER, 1997).

Já para análise múltipla de sequencias são utilizadas ferramentas com

características diferentes do BLAST, como é o caso do CLUSTALW, um programa para alinhamento múltiplo de sequências usado para a analise de proteínas e nucleotídeos. Segundo Thompson, Higgins e Gibson (1994) a ferramenta dispõe de um ambiente integrado para leitura de arquivos em vários formatos, alinhamento de sequências e analise de resultados. O método utilizado pelo CLUSTALW para o alinhamento múltiplo de sequências é o Alinhamento progressivo.

Ainda segundo Thompson, Higgins e Gibson (1994) esse método de alinhamento progressivo usam algoritmos de programação dinâmica[2] para calcular distâncias entre pares de sequências. As distâncias são usadas para construir uma árvore que serve de guia para criação do alinhamento múltiplo. Este método pode ser dividido em três passos nos quais cada passo gera informação necessária para o passo seguinte. O primeiro passo é responsável por gerar a matriz de distâncias entre as sequências de entrada, essa matriz contem valores percentuais relatando quão próximas estão duas sequências quaisquer de entrada. No segundo passo, é gerada a árvore guia para o último passo, pode ser vista como uma arvore filogenética[3], apesar de ser uma aproximação dela. No último passo é feito o alinhamento em si, usando a árvore do passo anterior como guia durante o processo.

### 2.4.3 Principais ferramentas para análises múltiplas de sequencias genômicas

Diversas ferramentas de Bioinformática foram desenvolvidas para viabilizar a análise simultânea de um grande número de sequencias genômicas, com o objetivo de viabilizar a realização de estudos filogenéticos, evolutivos e de comparações de mutantes. Esta forma de abordagem é comumente denominada análise múltipla de sequencias do inglês (MAS *Multiple Sequence Alignment)* (LIMA, 2012). A Tabela 1 descreve as principais ferramentas disponíveis livremente na internet para análises múltiplas de sequencias genômicas

[2] Programação dinâmica é um método para a construção de algoritmos para a resolução de problemas computacionais de otimização combinatória. É aplicável em problemas nos quais a solução ótima pode ser computada a partir da solução ótima previamente calculada e memorizada - de forma a evitar recálculo - de outros subproblemas que, sobrepostos, compõem o problema original.

[3] Arvore filogenética é uma representação gráfica, em forma de uma árvore, das relações evolutivas entre várias espécies ou outras entidades que possam ter um ancestral comum.

Tabela 1 – Descrição de *Softwares* para análises múltiplos de sequencias disponíveis gratuitamente na internet

| Software | Descrição |
|---|---|
| CLUSTALW | *Em bioinformática CLUSTAL é uma série de programas de computador amplamente utilizado para alinhamento múltiplos de sequencias: CLUSTAL, CLUSTALV, CLUSTALW, CLUSTALX E CLUSTAL OMEGA* |
| DIALIGN | *Série de programa de computador criado para alinhamento múltiplo de sequências. Variações: CHAOS-DIALIGN, DIALIGN-TX E DIALIGN-PFAM* |
| KALINGN | *Um algoritmo de alinhamento múltiplo rápido e preciso. É bastante utilizado para alinhar um grande número de sequencias.* |
| MAFFT | *MAFFT é um programa de alinhamento múltiplo de sequências para sistemas operacionais Unix-like. Ele oferece uma gama de vários métodos de alinhamento, L-INS-i (exato; para o alinhamento de aproximadamente 200 sequências), FFT-NS-2 (rápido, para o alinhamento de aproximadamente 30.000 sequências).* |
| MUSCLE | *comparação de sequência múltipla de Log-Expectativa. Dependendo das opções escolhidas pode alcançar melhor precisão e maior velocidade do que CLUSTALW e T-COFFEE.* |
| PCMA | *É um alinhador múltiplo baseado em consistência que faz uso da abordagem progressiva.* |
| PROBCONS | *PROBCONS é uma nova ferramenta para gerar alinhamentos múltiplos de sequências proteicas. Usando uma combinação de modelagem probabilística e técnicas de alinhamento baseadas em consistência.* |
| T-COFFEE | *Pode alinhar sequências de proteínas, assim como sequências de ADN / ARN.* |

Fonte: Lima (2012); Rigden e Mello (2012); Santos (2004); Ticona (2003); Debian (2014)
Nota: Dados Trabalhados pelo autor

### 2.4.4 A ferramenta BLAST

Segundo Altschul, Madden e Schäffer (1997), o BLAST é a ferramenta mais usada na busca de similaridade em sequências de proteínas e DNA, faz buscas em bancos de dados de sequências a partir de uma sequência consulta (*query*) do usuário, e tenta achar todas as sequências do banco (*subjects*) que têm alinhamentos estatisticamente semelhantes, neste caso, a questão principal se relaciona com falsos positivos e falsos negativos. Os falsos positivos significa resultado retornado pelo programa, mas que não corresponde a um resultado biologicamente verdadeiro e os falsos negativos é o resultado que o programa deixa de reportar, apesar de ser biologicamente verdadeiro.

O programa BLAST está baseado em heurística que traz uma melhora significativa nos tempos de resposta das buscas em bases de sequências. Faz comparação local usando uma subcadeia comum exata com um tamanho mínimo padrão de 11 bases consecutivas para DNA (SETUBAL; MEIDANIS, 1997).

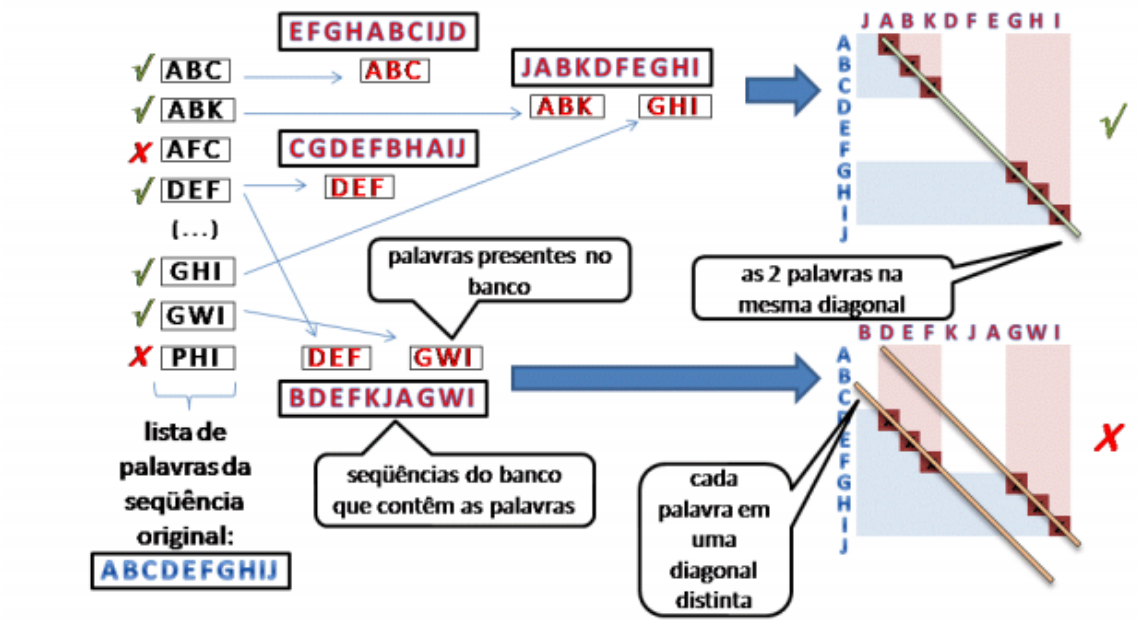
Segundo Gibas e Jambeck (2001), as comparações de sequências são feitas em pares, procurando sítios de similaridade local, ao invés de alinhamentos globais ótimos entre as sequências inteiras. O programa BLAST, que agiliza o alinhamento de sequência local, possui três etapas básicas. Primeiro, ele cria uma lista de todas as sequências curtas (chamadas palavras na terminologia do BLAST) com pontuação acima de um valor limite quando alinhadas com a sequência de pesquisa. Em seguida, o banco de dados é consultado para obter as ocorrências destas palavras. Como o comprimento da palavra é muito curto (três resíduos para proteínas, onze resíduos para ácido nucléico), é possível pesquisar uma tabela pré-calculada de todas as palavras e suas posições nas sequências, para obter um aumento de velocidade.

Estas palavras combinadas são estendidas para alinhamentos locais sem gaps entre a sequência de pesquisa e a sequência do banco de dados (RIGDEN; MELLO, 2002). Segundo estes autores, a pesquisa de similaridades de sequências dentro de um banco de dados é o primeiro passo na análise de uma nova sequência. Se a sua sequência desconhecida tiver uma cópia similar já descrita em um banco de dados, a busca irá fornecer rapidamente essa informação.

Na Figura 3 estão representados graficamente os procedimentos da ferramenta BLAST. A sequencia de entrada é dividida em palavras com tamanho 3.

Para cada palavra, resultante da divisão da sequencia de entrada, é realizada a consulta nas sequencias disponíveis no banco de dados. Para cada resultado positivo, a palavra encontrada é inserida numa matriz e a similaridade é identificada pelas palavra que ficarem posicionadas na mesma diagonal, conforme pode ser visto na Figura 3.

Figura 3 - Exemplo de procedimento do BLAST para buscas de sequencias proteicas , com tentativas inicias de todas as combinações possíveis de 3 aminoácidos consecutivos




Fonte: Amaral et al. (2007, p.9)

A pesquisa em banco de dados fornece os primeiros indícios de que sua sequência pertence a uma família bem conhecida de proteínas. Se há similaridade entre a sua sequência com as de outras espécies, então esta pode ser homóloga (ou seja, descende de uma sequência ancestral comum). Conhecer a função da sequência similar/ homóloga é uma boa indicação da identidade da sequência desconhecida.

Apesar de estar disponível em vários servidores Web, os usuários do BLAST podem fazer download e instalar o aplicativo em seus computadores. Uma vantagem em usar a versão web é a não preocupar-se em atualizar a ferramenta. Em caso de indisponibilidade de internet é possível instalar o BLAST e copiar uma base de dados para a realização de testes localmente e ainda inserir dados nesta

base, e criando o seu próprio banco de dados. Assim o usuário poderá ter informações que ainda não foram publicadas ou distribuídas (GIBAS; JAMBECK, 2001).

### 2.4.5 A Ferramenta CLUSTALW

Segundo Thompson, Higgins e Gibson (1994) o CLUSTALW é um programa para alinhamento múltiplo de sequências. Existem várias técnicas que podem ser utilizadas para resolução do problema de alinhamento múltiplo de sequências.

O Alinhamento Múltiplo Ótimo é caracterizado devido à complexidade de a solução ser muito custosa, problema NP-difícil[4], esta abordagem é muito pouco utilizada, pois somente resolve problemas com pouquíssimas sequências. O alinhamento progressivo é a abordagem utilizada pelo CLUSTALW, em que a solução não é ótima, mas gera um alinhamento com pouca perda de qualidade em tempo aceitável para uma quantidade razoável de sequências. E o alinhamento com heurísticas possui uma abordagem utilizada quando se quer alinhar uma grande quantidade de sequências de uma base de dados (THOMPSON; HIGGINS; GIBSON 1994).

Ainda de acordo com Thompson, Higgins e Gibson (1994) o alinhamento múltiplo pode ser usado em situações que se deseja encontrar padrões que caracterizam famílias de sequências, detectar ou demonstrar homologia entre novas sequências, ajudar a predizer as estruturas secundárias e terciárias de novas sequências, sugerir oligonucleotideos primários para PCR e analise da evolução nuclear.

O programa pode ser usado para alinhamento múltiplo de sequências de proteínas e nucleotídeos. A ferramenta dispõe de um ambiente integrado para leitura de arquivos em vários formatos (THOMPSON; HIGGINS; GIBSON, 1994).

A ferramenta está disponível livremente na internet em http://www-igbmc.u-strasbg.fr/BioInfo/, onde também é encontrado o CLUSTALX, versão baseada no CLUSTALW com interface gráfica (com janelas, menus e ajuda).

---

[4] Problema NP-Difícil constitui-se de todos os problemas que não possuem solução determinística. Não necessariamente possuem uma forma eficiente de verificar suas soluções. Não necessariamente existem métodos polinomiais de verificar se uma resposta para um problema de decisão dessa classe é correta.

### 2.4.6 A Ferramenta T-COFFEE

Segundo Debian (2006) T-COFFEE é um pacote de softwares de alinhamentos de sequências múltiplas. A partir de um conjunto de sequências (proteínas ou DNA) de entrada, o T-COFFEE gera um alinhamento de sequências múltiplas. A versão 2.00 e superiores podem misturar sequências de aminoácidos com sequencia de nucleotídeos. O T-COFFEE permite a combinação de uma coleção de alinhamentos múltiplos/pareados globais ou locais para um modelo único. Esta ferramenta também permite estimar o nível de consistência de cada posição dentro do novo alinhamento com o resto dos alinhamentos.

No alinhamento com nível de consistência são incorporadas informações das diferentes sequencias em cada criação de pares de alinhamento. Na primeira etapa de um alinhamento progressivo é convencional utilizar apenas informações de sequencias para cada dois alinhamentos pareados. Esta estratégia tende a gerar alinhamentos de sequencias muito mais precisa. O T-COFFEE é uma dos algoritmos mais conhecidos.

Debian (2006) afirma que o T-COFFEE tem uma chamada especial ao M-COFFEE que torna possível combinar o resultado de muitos pacotes de alinhamento de sequências múltiplas tais como: MUSCLE, PROBCONS, POA, DiAlign-TS, MAFFT, CLUSTAL W, PCMA e T-COFFEE.

Esta ferramenta pode ser instalada em um computador ou pode ser utilizada a partir do navegador Web, onde são inseridos os dados em um formulário on-line. Inicialmente o usuário define a entrada de dados na ferramenta (por exemplo, sequências, bases de dados). Nas etapas seguintes, o usuário tem a possibilidade de alterar os parâmetros de configuração da ferramenta. E, finalmente, o último passo é sempre o passo de submissão da ferramenta, onde o usuário pode especificar um título para ser associado com os resultados e um endereço de e-mail para notificação.

Após a validação de entrada dos dados é iniciado o processo de comparações. Ao final do processamento a página é direcionado para um link com os resultados e o mesmo link é enviado para o e-mail solicitado no formulário. Os parâmetros são validados antes de a ferramenta lançar os dados no servidor, em caso de erro de parâmetros, o usuário será notificado diretamente no formulário. (EBI 2014) http://www.ebi.ac.uk/Tools/msa/tcoffee/help/. Na sessão 4.1.3 encontram-

se mais detalhes de como usar a ferramenta.

### 2.4.7 2.4.7 A ferramenta MAFFT

MAFFT (*Multiple Alignment using Fast Fourier Transform)* é um programa para alinhamento de sequências múltiplas de alta velocidade. Utilizar a ferramenta no ambiente web é bastante semelhante ao uso do T-COFFEE. Em um formulário são inseridas as sequências ou a base de dados de entrada. Na etapa seguinte o usuário tem a opção de alterar os parâmetros padrões da ferramenta e finalmente submeter os dados às análises. Também pode ser especificado um título para ser associado com os resultados e um e-mail para onde será enviado o link com os resultados. Estas informações estão contidas na documentação da ferramenta MAFFT, disponível em: http://www.ebi.ac.uk/Tools/msa/mafft/help/.

MAFFT, T-COFFEE, CLUSTALW E BLAST não são as únicas ferramentas disponíveis em ambiente web. Várias outras podem ser acessadas ou instaladas usando sistemas operacionais diferentes, tais como Windows, Linux e MAC OS.

A Web também possibilita que bancos de dados públicos de sequências de genoma ofereçam serviços por meio de uma interface uniforme para uma comunidade mundial de pesquisadores. Com um programa aplicativo comum, um biólogo molecular pode, então, comparar uma sequência de DNA desconhecida com a coleção completa de sequências de DNA públicas.

## 2.5 Bancos de dados de sequências de DNA

Pode se considerar um banco de dados uma coleção de dados inter-relacionados projetados para organizar e estruturar dados para facilitar o acesso aos dados. O primeiro banco de dados de sequências de DNA, instalado foi o GSDB (*Gene Sequence Database*) em 1979, no Laboratório Nacional de Los Alamos. Em 1988 o governo americano lançou o primeiro banco de dados público contendo sequências de DNA dos mais diversos organismos. Este repositório de sequências instalou-se através do Centro Nacional de Informações de Biotecnologia - NCBI e o banco de dados passou a ser conhecido por *GenBank* (BENSON, 2010).

Na primeira década de sua construção o *GenBank* cresceu timidamente, mas com o advento do projeto Genoma Humano e os avanços

tecnológicos o seu crescimento foi acelerado. Segundo Gibas e Jambeck (2001) especialistas em computação tem trabalhado junto com pesquisadores das ciências biológicas para armazenar e processar um enorme volume de informação da melhor forma possível.

Sansom e Smith (2000) afirmam que o projeto Genoma Humano contribuiu significativamente para o crescimento da bioinformática nas décadas e 80 e de 90. A quantidade crescente de informações juntamente com o incremento da sofisticação das técnicas de clonagem e sequenciamento levou a criação de softwares destinados a esse processo e também a redução dos custos operacionais.

Hoje já são dezenas de banco de dados com bilhões de informações sobre genes das mais variadas espécies e grande parte deles possui interface para a web. Gibas e Jambeck (2001) afirmam que os usuários determinam suas ferramentas de pesquisas e o banco de dados conforme os objetivos da pesquisa. O banco de dados com maior número de informações e atualizações é o *GenBank* do NCBI.

Existem diversos bancos de dados com informações sobre genes e suas mutações. Os principais são: banco de dados dos Estados Unidos, mantido pelo NCBI (*National Center for Biotechnology Information*); banco de dados da Europa mantido pelo EBI (Instituto Europeu de Bioinformática) e o banco de dados de DNA do Japão (DDBJ). Esses bancos de dados são importantes fontes de informações sobre mutações utilizadas por vários cientistas para conduzirem suas pesquisas, reduzindo significativamente o tempo necessário para levantarem dados manualmente.

### 2.5.1 Análise de mutações (triagem de mutantes)

Mutações no gene supressor de tumor, p53, são as alterações genéticas mais comuns encontradas em câncer humano. O tipo e localização das mutações no gene são diferentes em distintos cânceres humanos. Esta variabilidade e espectro de mutações podem representar fatores etiológicos agindo no processo de transformação maligna (SIMÃO et al. 2002).

As mutações encontradas no gene p53 podem ser somáticas, ocorrendo no tecido tumoral ou germinativa e podem ser encontradas em todas as células do organismo, incluindo células reprodutivas, o que caracteriza a Síndrome de Li-Fraumeni. Mutações somáticas encontradas entre os cânceres humanos que

foram identificados por sequenciamento, apresentam 87,2% de substituições de uma única base e 12,8% apresentam mutações mais complexas, deleções e inserções. As mutações que promovem a troca de 16 aminoácidos (missense) são observadas em 231 dos 393 códons (HAINAUT; HOLLSTEIN, 2000), enquanto que as mutações silenciosas representam 3,9% das mutações descritas (STRAUSS, 1997). De todas as mutações já identificadas, 93% estão presentes no domínio de ligação ao DNA e correspondem aos exons 5,6,7,e 8 do gene p53. As inserções, deleções e códons de terminação representam 48% das mutações no domínio N-terminal e 60% no C-terminal, mas apenas 17% no domínio de ligação ao DNA (HAINAUT; HOLLSTEIN, 2000).

Como já mencionado, mutações no gene TP53 não ocorrem apenas a nível somático, mas também em linhagens germinativas. A detecção dessas mutações germinativas é importante na identificação de indivíduos com alto risco de desenvolver câncer (FREBOURG et al., 1992). Mutações germinativas em TP53 predispõem uma variedade de tumores malignos de desenvolvimento precoce incluindo câncer de mama, câncer cerebral, sarcoma ósseo e carcinoma adrenocortical (VARLEY et al. 2001). Essas mutações foram inicialmente relatadas em pacientes com a Síndrome de Li-Fraumeni, uma síndrome rara de câncer familiar, onde parentes afetados desenvolvem diversos tumores malignos (FREBOURG et al., 1992; KLEIHUES et al., 1997).

De maneira geral, o câncer é uma doença causada por mutações genéticas que conferem às células algumas características especiais, como capacidade ilimitada de proliferação, perda de resposta a fatores de inibição de crescimento, evasão de apoptose (morte celular programada), capacidade de invadir outros tecidos (metástases) e produção de novos vasos sanguíneos (angiogênese).

### 2.5.2 O Câncer de pele

A maioria dos tumores malignos cutâneos não herdados resulta de mutações causadas por carcinógenos que, de alguma forma, promovem danos ao DNA e conferem vantagens que propiciam o crescimento celular desordenado e a invasão de outros tecidos. Os fatores de risco que contribuem para o desenvolvimento do câncer de pele são bem conhecidos e inclui principalmente raça, idade, gênero, exposição crônica a agentes mutagênicos químicos e físicos,

além de fatores genéticos (MARTINEZ et al., 2006).

De acordo com o INCA (2014) o câncer de pele não melanoma é o tipo mais frequente no Brasil e corresponde a 25% de todos os tumores malignos registrados no país. Se for detectado precocemente, tem alto percentual de cura. Entre os tumores de pele o tipo não melanoma é o que tem maior incidência e menor mortalidade.

Thompson (1994) ressalta que o processo de formação de tumores começa apenas por uma célula somática em algum lugar do organismo.

### 2.5.3 O Gene p53 e o câncer de pele

O gene p53 é um gene inibidor do ciclo celular e no momento em que ocorrem danos ao DNA são acionados mecanismos bioquímicos capazes de reparar tais lesões. Para isso é necessário que a célula interrompa o ciclo celular para que a mutação não se perpetue. Caso ocorra falha nesse processo de reparo, a proteína força a apoptose que é a morte programada da célula. Mutações no gene p53 estão entre as alterações mais frequentes encontradas em diversos tipos de canceres. Quando a proteína sofre mutações se torna incapaz de impedir a divisão celular, permitindo a proliferação de células com erros, cujo acúmulo pode levar à ativação de oncogêneses, ou perda da função dos genes supressores de tumor. Também podem inibir a apoptose e assim aumentar a sobrevida de células alteradas. (MARTINEZ et al. 2006).

Magnoni e Francisco (2009) afirmam que sexo e idade não estão relacionados à expressão da proteína P53, mas as substâncias contidas no cigarro influenciam na expressão da proteína e também fatores ambientais como exposição excessiva aos raios ultravioletas.

Segundo Corapcioglu et al. (2006) o gene p53 é um dos genes mais estudado pela comunidade científica mundial. Está localizado no braço curto do cromossomo 17. Sua mutação está presente em aproximadamente 50% de todos os cânceres humanos e representa a alteração genética mais comum em células malignas (MORITA; IKEDA; TAKAMI, 2008). O dano no DNA foi o primeiro tipo de estresse descoberto na ativação do p53 e, por isso, esse gene é amplamente considerado como o “guardião do genoma humano” (EFEYAN; SERRANO, 2007).

Na presente pesquisa limitaram-se as sequencias do gene p53

relacionadas aos canceres cutâneos de pessoas localizadas em países com diferenças contrastantes. Mas pesquisadores de diversos países tem se empenhado em pesquisar os genes dos mais variados tipos de canceres.

### 2.5.4 Estudos correlacionados com o gene p53

Em uma ampla revisão bibliográfica realizada durante a presente pesquisa foram encontrados trabalhos em que a pesquisa tinha foco especificamente sobre os variados tipos de câncer de pele (PARKIN, 2004; CORAPCIOGLUD et al., 2006; MARTINEZ et al., 2006; EFEYAN; SERRANO, 2007), outros trabalhos com foco no gene p53 (FREBOURG et al., 1992; KLEIHUES et al., 1997; STRAUSS, 1997; HAINAUT; HOLLSTEIN, 2000; VARLEY et al., 2001; SIMÃO, 2002; MORITA; IKEDA; TAKAMI, 2008; MAGNONI; FRANCISCO, 2009) e, finalmente, outros relatos sobre a utilização de algoritmos para análise múltipla de sequencias (THOMPSON; HIGGINS; GIBSON, 1994; ZHANG; MASSEN, 1997; SANSOM; SMITH, 2000; PUCCI NETO, 2001; PROSDOCIMI et al. 2002; RIGDEN; MELLO, 2002; TICONA, 2003; SANTOS, 2004; OLIVEIRA, 2008; LIMA, 2012; DEBIAN, 2014).

Porém não foram localizados trabalhos que, especificamente, explorasse o assunto relacionando análises de sequencias de nucleotídeos extraídas de pacientes humanos com diagnóstico comprovado de câncer de pele, utilizando softwares de análise múltipla de sequencias empregados na presente pesquisa.

# 3 MATERIAIS E METODOLOGIAS

## 3.1 Materiais

**Equipamentos:** Foram utilizados um computador com Processador Intel Core i5, 4Gb de memória, HD de 1Tb com placa de vídeo 1GB, Monitor de 20" e configurado com permissão de administrador; Impressora jato de tinta colorido; Scanner de mesa e acesso irrestrito a internet. Os equipamentos foram instalados no Laboratório de Citogenômica e Bioinformática da UNOESTE, com a finalidade de gerenciar, processar e armazenar as informações de materiais fornecidos por pesquisadores, e assim criar bases de dados próprias.

**Bancos de dados:** Foram pesquisadas sequências de nucleotídeos correspondentes ao gene p53, relacionados com a ocorrência de Câncer de pele em humanos, localizadas nas bases de dados de acesso público na internet. Para a presente pesquisa foram escolhidos os bancos de dados dos institutos americano (denominado NCBI – *National Center for Biotechnology Information*), europeu (EBI – *European Bioinformatics Institute*) e japonês (DDBJ – *DNA Databank of Japan*), descritos adiante.

**Sequências nucleotídicas** do gene p53 foram comparadas, através de ferramentas de análise múltiplas de sequencias (CLUSTALW, MAFFT e T-COFFEE), com as sequências de variantes genéticos (mutantes) descritos em pacientes acometidos com câncer de pele. Devido ao fato destas sequencias estarem armazenadas em bancos genômicos, de acesso público, localizados em diversos países, supõe-se que isto implique em diferentes situações de impacto ambiental e com diferenças contrastantes em relação aos fatores: grupo étnico, localização geográfica em respeito à latitude e valores de IDH, entre outros.

## 3.2 Metodologia de coleta de dados em banco de sequencias

A escolha do gene a ser pesquisado foi o passo inicial desta pesquisa. Segundo Martinez et al. (2006) o gene p53 é um gene inibidor do ciclo celular e no momento em que ocorrem danos ao DNA são acionados mecanismos bioquímicos capazes de reparar tais lesões. Para isso é necessário que a célula interrompa o ciclo celular para que a mutação não se perpetue. Caso ocorra falha nesse processo de reparo, a proteína força a apoptose que é a morte programada da célula.

Mutações no gene p53 estão estre as alterações mais frequentes encontradas em diversos tipos de canceres. Quando a proteína sofre mutações se torna incapaz de impedir a divisão celular, permitindo a proliferação de células com erros, cujo acúmulo pode levar à ativação de oncogêneses ou perda da função dos genes supressores de tumor. Martinez et al. (2006) também afirmam que os raios ultravioletas, especificamente os ultravioletas tipo B, são fonte de mutações para o referido gene.

Após a identificação do gene foram escolhidos os bancos de dados onde seriam pesquisadas as sequencias de aminoácidos do gene p53. Vários institutos de pesquisas disponibilizam gratuitamente o acesso aos seus bancos de dados, com milhares de sequencias de genes, das mais variadas espécies.

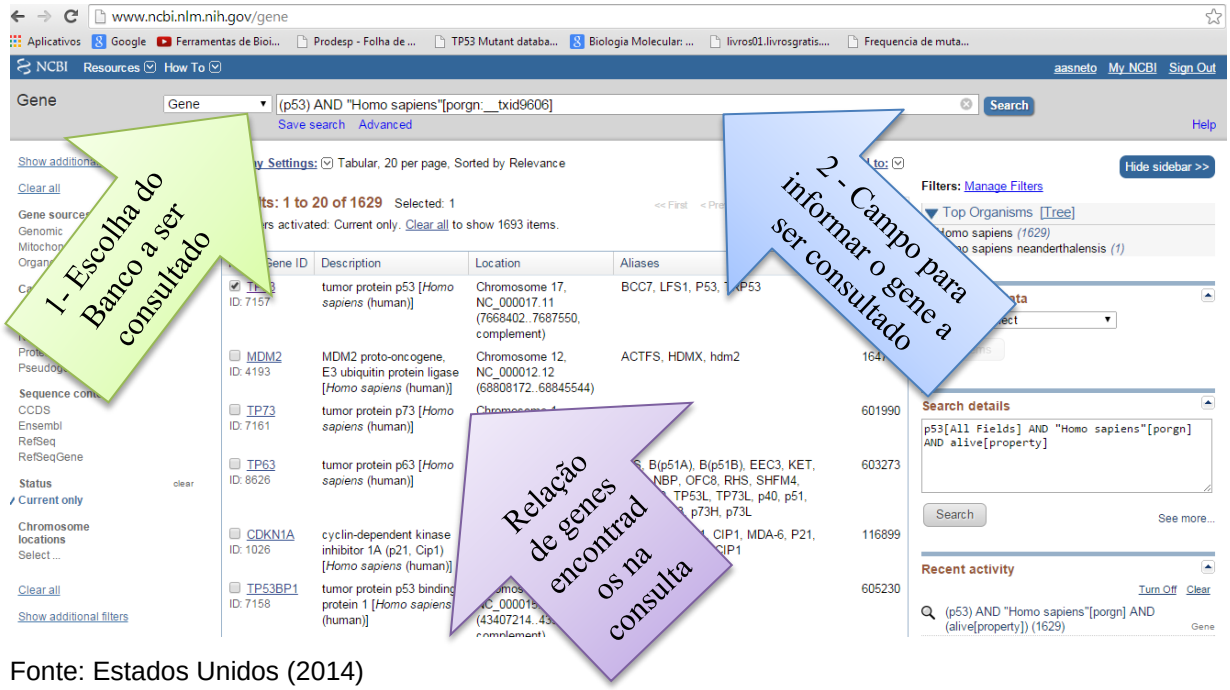
### 3.2.1 NCBI

O Centro Nacional de Informações sobre Biotecnologia (NCBI - http://www.ncbi.nlm.nih.gov/) está localizado nos Estados Unidos da América e é reconhecido mundialmente com uma das referencias mundiais em informações sobre biotecnologia. Possui como visão encontrar novas abordagens para lidar com um grande volume de dados complexos disponibilizando aos pesquisadores acesso e ferramentas de análises computacionais para avançar na compreensão da genética e seu papel na saúde e na doença. O conteúdo do site é de acesso público e várias ferramentas de análises estão disponíveis tanto para download como para funcionarem no ambiente web, ou seja, sem que haja a necessidade do software ser instalado no computador do pesquisador, evitando assim a necessidade de averiguar a existência de atualizações.

Dentre os serviços disponibilizados no site pode-se destacar a possibilidade de realizar consultas sobre genes, proteínas e sequências de DNA das mais variadas espécies aplicando filtros de refinamento na pesquisa, descritos a seguir. Na maioria dos resultados estão disponibilizados os links das publicações, referenciando seus autores e títulos dos trabalhos concluídos.

A busca foi iniciada pelas mutações do gene p53 no banco de dados do NCBI para tanto, foram obedecidos alguns parâmetros conforme identificado na Figura 4.

Figura 4 – Página do instituto americano de bioinformática (NCBI) com as opções de busca de gene



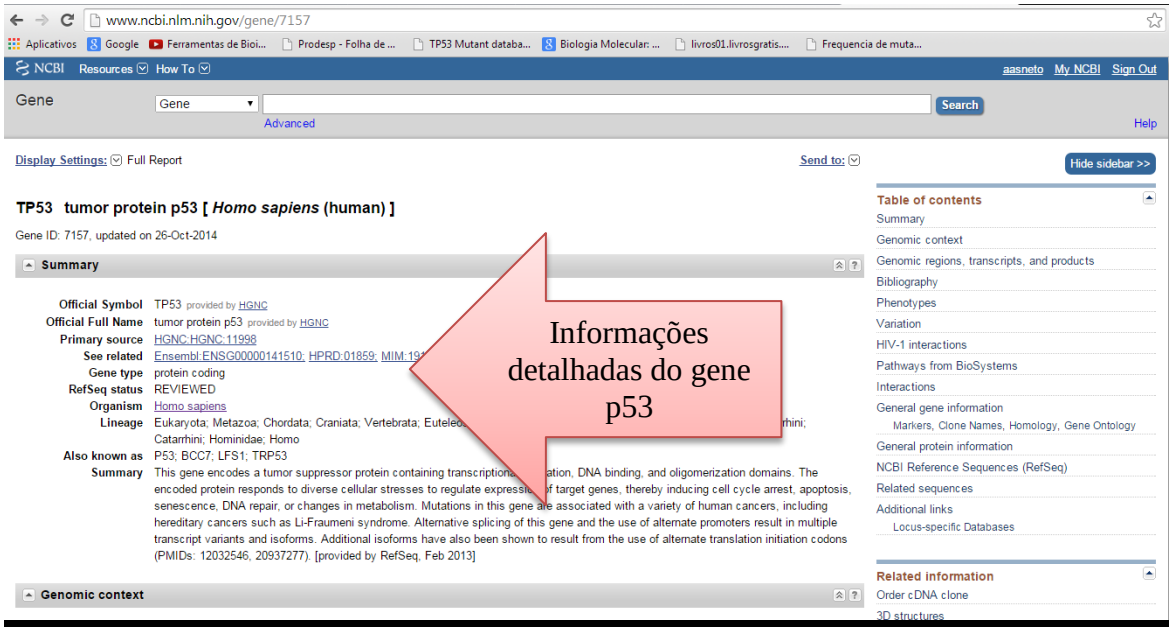

Fonte: Estados Unidos (2014)

OBS: Dados trabalhados pelo autor

A seta número 1 identificado na Figura 4 aponta para o campo de escolha do banco de dados a ser consultado. A seta número 2 aponta para o campo onde deve ser digitado os parâmetro a ser localizado “gene p53 e humanos” e a seta número 3 aponta para a relação de genes encontrados.

Após a identificação do gene, conforme indicado na seta número 3, na Figura 4, o passo seguinte é a exibição das informações detalhada do gene, tais como o símbolo oficial, nome completo do gene, tipo do gene, o organismo, linhagem, um sumário e a representação gráfica da localização do gene dentro do cromossomo vistos na Figura 5 e Figura 6.

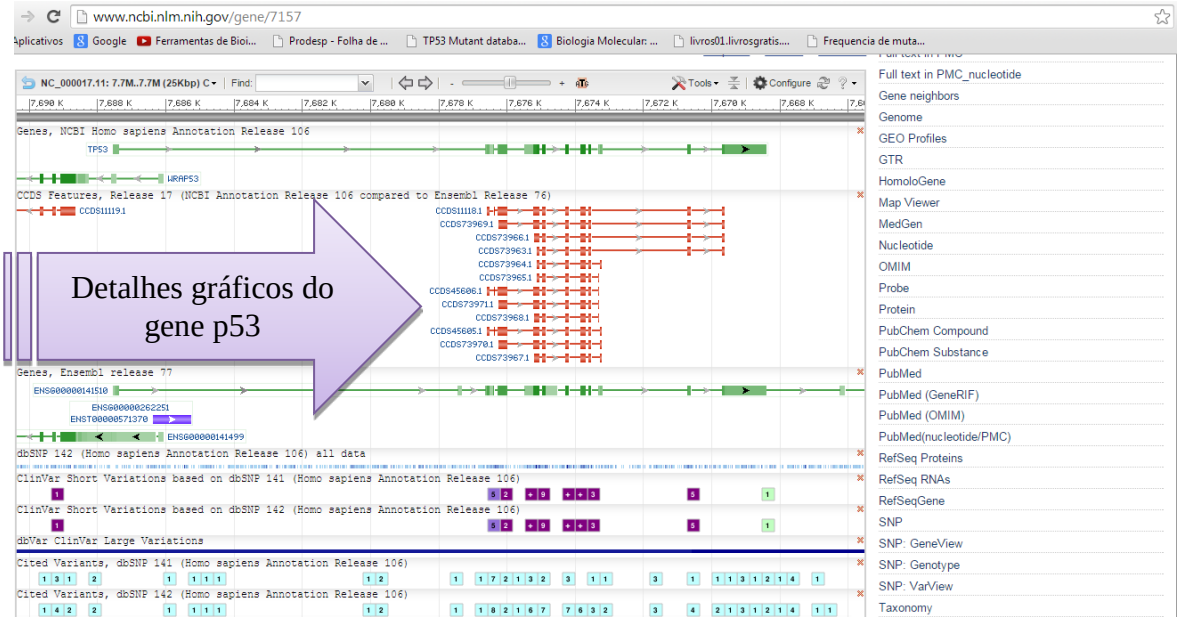
Figura 5 - Página do instituto americano de bioinformática (NCBI) com informações detalhada do gene p53




Fonte: Estados Unidos (2014)

OBS: Dados trabalhados pelo autor

Figura 6 - Informações Gráficas do gene p53 visualizadas na página do instituto americano de bioinformática (NCBI)



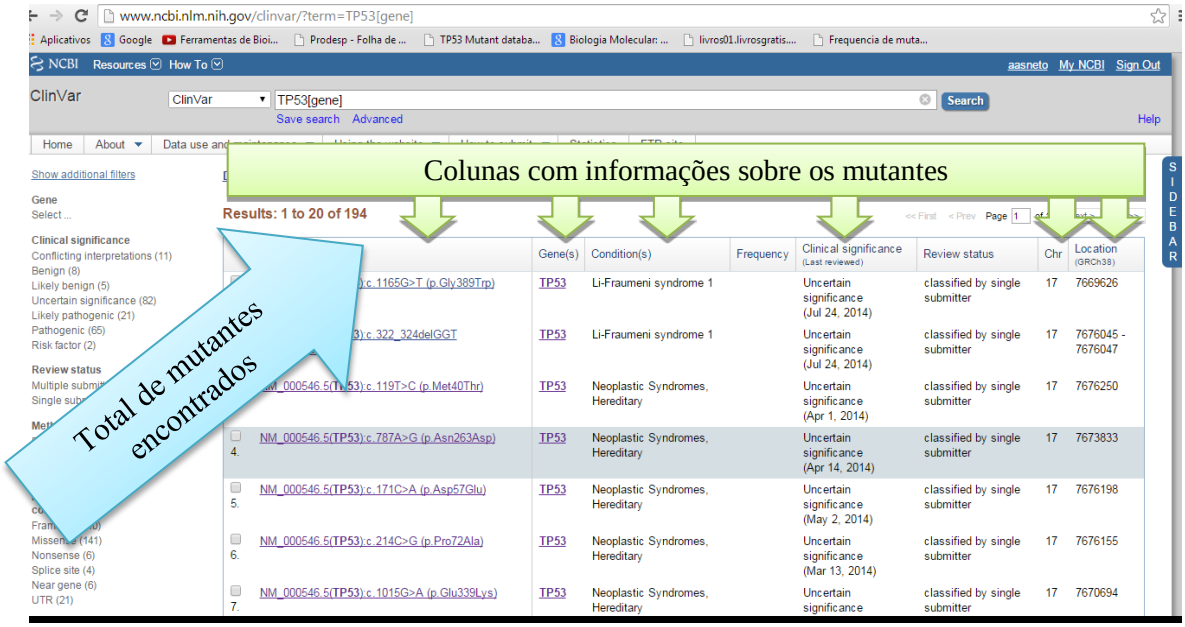

Fonte: Estados Unidos (2014)

OBS: Dados trabalhados pelo autor

Abaixo das informações gráficas do gene está o link referente a uma lista com as variações encontradas para este gene armazenados em uma base de

dados indicada como *ClinVar*. Cada linha da lista refere-se a um mutante e as colunas exibem informações sobre o mutante.

Figura 7 – Relação de Mutantes encontrados na página do instituto americano de bioinformática (NCBI)



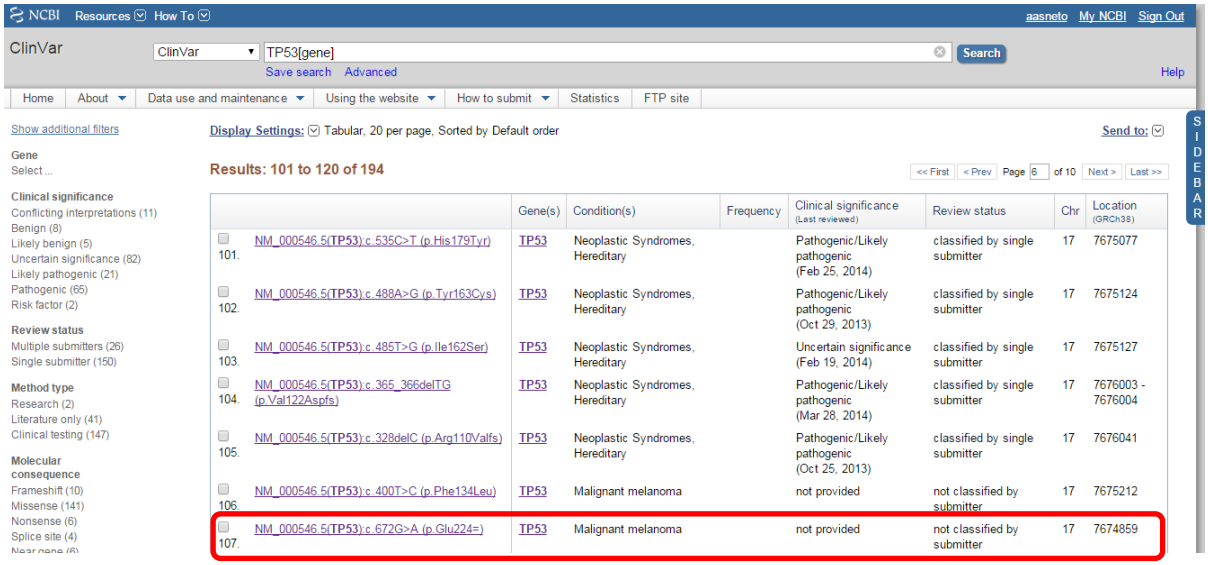

Fonte: Estados Unidos (2014)

OBS: Dados trabalhados pelo autor

A lista apresentada no formato de uma tabela exibe algumas informações sobre a mutação, tais como o nome do mutante, o gene, a patologia que se refere à mutação, o nome do laboratório e a data em que foi sequenciado, o número do cromossomo e a localização dentro do cromossomo. Algumas dessas informações no formato de *link* podem ser expandidas ao clique do mouse.

O número de mutações encontrados em outubro de 2014 para o gene p53 no banco de dados do NCBI foram de cento e noventa e quatro. Para a presente pesquisa, foram consideradas apenas as mutações identificadas como provenientes de material sequenciado de pacientes com canceres cutâneos.

Figura 8 – Ilustração da Identificação do gene mutante encontrado na página do Instituto Americano de Bioinformática (NCBI)




Fonte: Estados Unidos (2014)

OBS: Dados trabalhados pelo autor

Do total de 194 mutações, foram copiadas as sequencias de aminoácidos de 6 variações e do gene p53 do banco de dados do Instituto Americano e armazenadas em um arquivo texto para posteriormente serem processadas com maior facilidade.

Cada mutação na tabela refere-se a um link com dezenas de informações sobre o gene mutante das quais se destacam várias referencias de publicações sobre a mutação, a sequencia de aminoácidos, a sequencia de nucleotídeos a localização do laboratório ou instituto onde o material genético foi sequenciado, a descrição da mutação indicando o códon onde ocorre a mutação e o tipo de alteração: mudança, inserção ou deleção.

Para a presente pesquisa escolheu-se o formato *FASTA,* reconhecido e aceito como formato de entrada em vários softwares de análises de sequencias.

O formato *FASTA* é uma sequência de caracteres que representam os nucleotídeos ou os aminoácidos. Como regra é inicializado com o sinal ‘>’ e a primeira linha é reservada para as informações da sequência, tais como nome ou código da sequência, número do cromossomo. A primeira linha, que pode ser considerada como cabeçalho da sequência, não tem um padrão oficialmente definido de quais informações devem constar sobre as sequências. As informações que constam na primeira linha podem variar de um banco de dados para outro.

Figura 9 - Formato FASTA da sequencia do gene p53 localizado na página do instituto americano de bioinformática (NCBI)

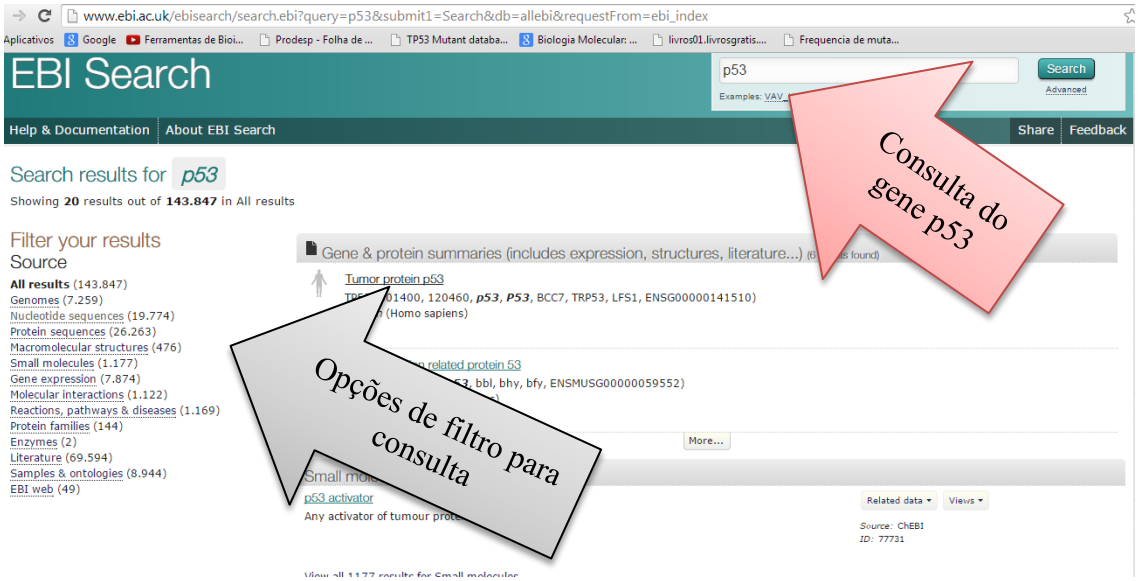

Fonte: Estados Unidos (2014)

### 3.2.2 EBI

O *European Bioinformatics Institute* (EBI - http://www.ebi.ac.uk/) é um complexo multicêntrico que agrega diversos países da Europa Ocidental, e é reconhecido mundialmente com uma das referencias mundiais em informações sobre bioinformática. Possui como visão encontrar novas abordagens para lidar com um grande volume de dados complexos disponibilizando aos pesquisadores acesso e ferramentas de análises computacionais para avançar na compreensão da genética e seu papel na saúde e nas principais doenças humanas. O conteúdo do site também é de acesso público e várias ferramentas de análises estão disponíveis aos usuários, tanto para download como para funcionarem no ambiente web, ou seja, sem que haja a necessidade do software ser instalado no computador do pesquisador, evitando assim a necessidade de sempre averiguar se existe atualização.

Para localizar os mutantes do gene p53 disponibilizados no site, foi feito uma consulta pelo gene p53 no espaço reservado para realizar buscas, conforme demonstrado na Figura 10.

Figura 10 - Identificação de parâmetros de busca do gene p53 na página Oficial do instituto europeu de bioinformática EBI



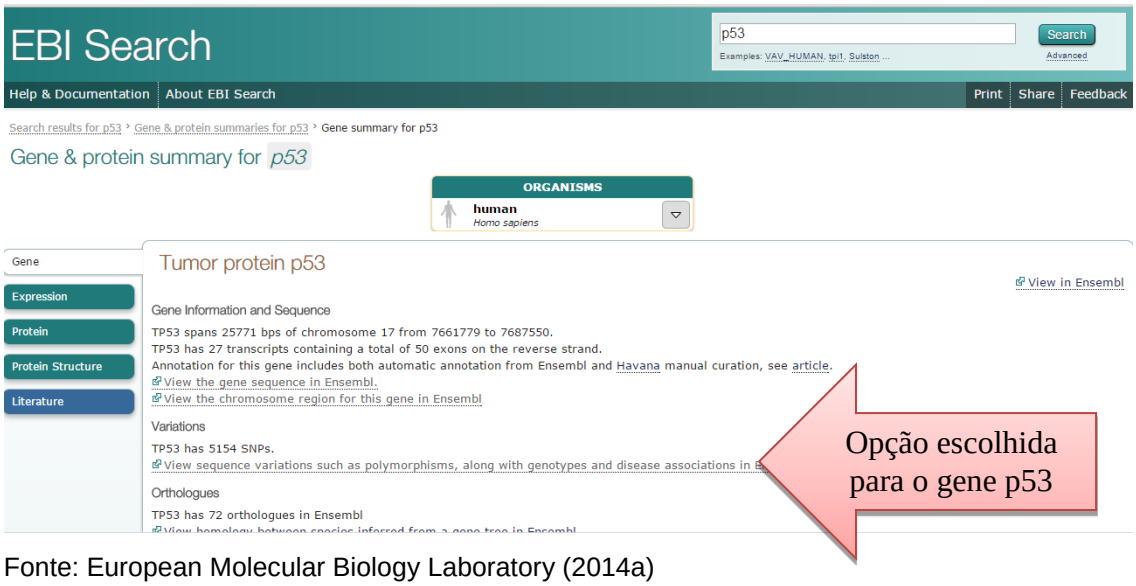

Fonte: European Molecular Biology Laboratory (2014a)

OBS: Dados trabalhados pelo autor

A Figura 10 apresenta o retorno da consulta realizada sobre o gene p53 no site oficial do EBI. Nesta página é possível aplicar vários filtros na consulta e refinar a busca. A página apresenta várias sugestões para refinamento da consulta ou até mesmo atalhos dos serviços mais utilizados pelos pesquisadores em consultas aos genes. Como a figura é apenas um recorte da página, várias opções só são visíveis se estiver online e utilizar a barra de rolagem do navegador.

Como esta pesquisa limitou-se em apenas a variações do gene p53 em humanos, com câncer de pele, foi escolhido o link "*Tumor protein p53*" localizado no inicio da página e identificado com o desenho de um corpo humano, conforme visto na Figura 10.

Figura 11 - Opções de filtros de busca do gene p53 na página Oficial do instituto europeu de bioinformática EBI

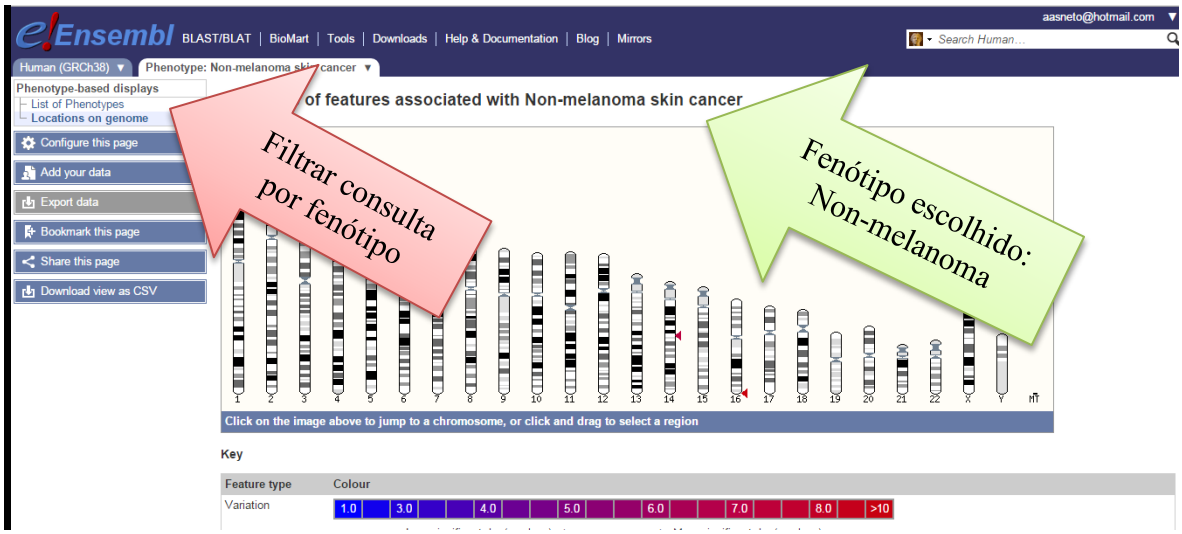



Fonte: European Molecular Biology Laboratory (2014a)

OBS: Dados trabalhados pelo autor

A Figura 11 exibe mais opções de filtros para a consulta do gene. Para esta pesquisa foi escolhida a opção *variations* e na sequencia escolhemos os fenótipos, ou seja, as variações das sequencias de nucleotídeos provenientes de pacientes diagnosticados com algum tipo de câncer, conforme visto na Figura 12.

Figura 12 – Detalhe gráfico apontando a localização do gene mutante associado com o fenótipo não melanoma encontrado na página Oficial do instituto europeu de bioinformática EBI

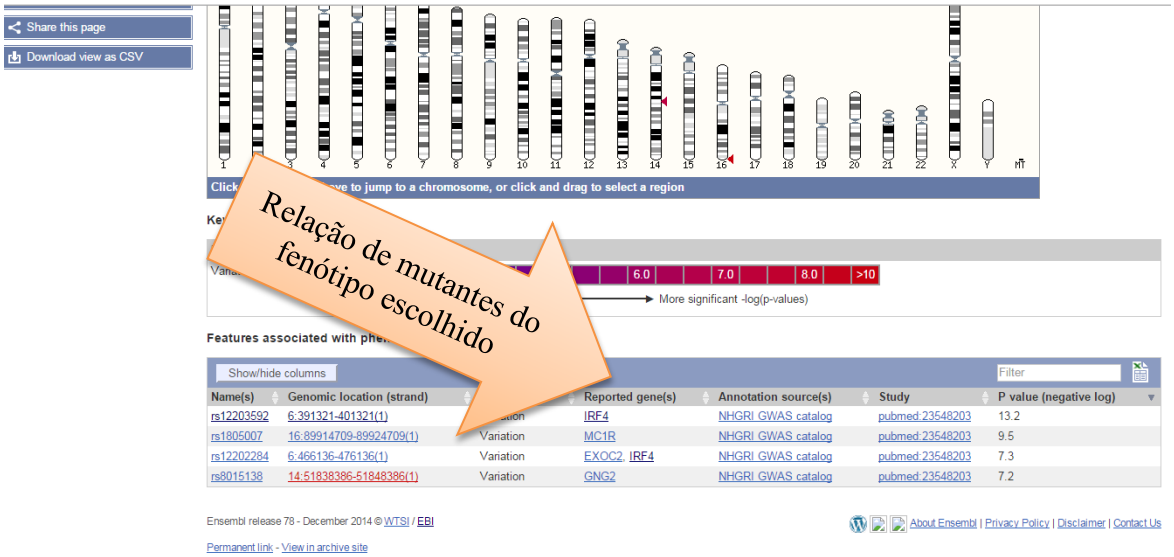



Fonte: European Molecular Biology Laboratory (2014a)

OBS: Dados trabalhados pelo autor

A Figura 13 exibe uma relação com quatro mutações do gene p53 para o câncer de pele do tipo não melanoma encontradas no banco de dados do EBI. Nesta pesquisa também foram encontradas mutações para outras variedades de câncer de pele.

Figura 13 - Lista de Mutantes encontrados na página Oficial do instituto europeu de bioinformática EBI para o fenótipo não melanoma




Fonte: European Molecular Biology Laboratory (2014a)

OBS: Dados trabalhados pelo autor

Na Figura 14 é possível visualizar a relação de mutantes com algumas informações sobre as mutações encontradas para o fenótipo identificado acima da relação. Cada link da tabela reporta para características detalhadas sobre a mutação, tais como, publicações científicas, anotações, localização no genoma e o nome ou identificação da mutação. Para cada mutante foi selecionado apenas a região da sequencia mutante, disponível no link do nome do mutante.

Figura 14 - Informações detalhadas do mutante encontrado na página Oficial do instituto europeu de bioinformática EBI


Fonte: European Molecular Biology Laboratory (2014a)

OBS: Dados trabalhados pelo autor

A sequência está disponível no link "*Flaking sequence*" localizado na lateral esquerda da página, conforme visto na Figura 14. Para cada mutante, foram identificadas as sequencias e copiadas no formato FASTA, para um arquivo do tipo texto. A Figura 15 exibe a tela com a opção de download da sequencia.

Figura 15 - Página Oficial do instituto europeu de bioinformática EBI com a opção de cópia da sequencia mutante encontrada


Fonte: European Molecular Biology Laboratory (2014a)

OBS: Dados trabalhados pelo autor

### 3.2.3 DDBJ

O *DNA Databank of Japan* (DDBJ) está localizado no Japão e é reconhecido mundialmente como um dos maiores repositórios de sequencias gênicas do mundo. No sitio está descrito que o instituto possui como missão encontrar novas abordagens para lidar com um grande volume de dados complexos disponibilizando aos pesquisadores acesso e ferramentas de análises computacionais para a análise de relações filogenéticas entre as diferentes sequências depositadas. O conteúdo do site é de acesso público e várias ferramentas de análises estão disponíveis tanto para download como para funcionarem no ambiente web.

# 4 RESULTADOS

Durante esta pesquisa foram realizadas várias consultas aos bancos de dados do Instituto Americano NCBI, europeu EBI e japonês DDBJ, referente ao gene p53. A sequencia original do gene foi localizada no NCBI e armazenada em um arquivo do tipo texto. Também foram pesquisadas as mutações referentes ao gene p53 e foram encontradas no banco de dados do NCBI cento e noventa e quatro mutações relacionadas a diversos tipos de canceres. Dentre estas foi selecionado seis sequencias de aminoácidos da região mutante do gene p53, identificadas como proveniente de material biológico sequenciado de pessoas com câncer de pele. As sequências foram copiadas e armazenadas no mesmo arquivo texto da sequencia original.

No banco de dados do instituto europeu foi possível consultar as mutações por fenótipos. Foram encontradas sessenta e duas sequencias de aminoácidos da região mutante do gene p53, identificadas como proveniente de material biológico sequenciado de pessoas com variados tipos de canceres. Para armazenar as sequencias mutantes encontradas no EBI foi criado um novo arquivo texto contendo a sequencia original do gene p53 e as sequencias variadas foram inseridas logo abaixo da original.

Na base de dados do Japão foram localizados quatro mil e setenta e cinco mutantes, mas não foi possível classificá-las como provenientes de pacientes com câncer de pele porque no momento da pesquisa não constava tal informação. Portanto, não foram utilizadas sequencias do Japão nas ferramentas de análise múltiplas.

## 4.1 Utilização de ferramentas para a comparação de sequências gênicas

As sequencias copiadas foram armazenadas em arquivos textos. Um arquivo texto para as sequencias encontradas nos bancos de dados do NCBI e outro arquivo texto para as sequencias encontradas no banco de dados do EBI e um arquivo também do tipo texto reunindo as sequencias dos dois bancos de dados.

Foi realizado um levantamento para identificar qual ferramenta seria utilizada para a análise de sequencias múltiplo e foi encontrada uma variedade significativa conforme detalhado na Tabela 1. Foram consideradas apenas ferramentas armazenadas ou disponibilizadas por centros de pesquisas de

bioinformática.

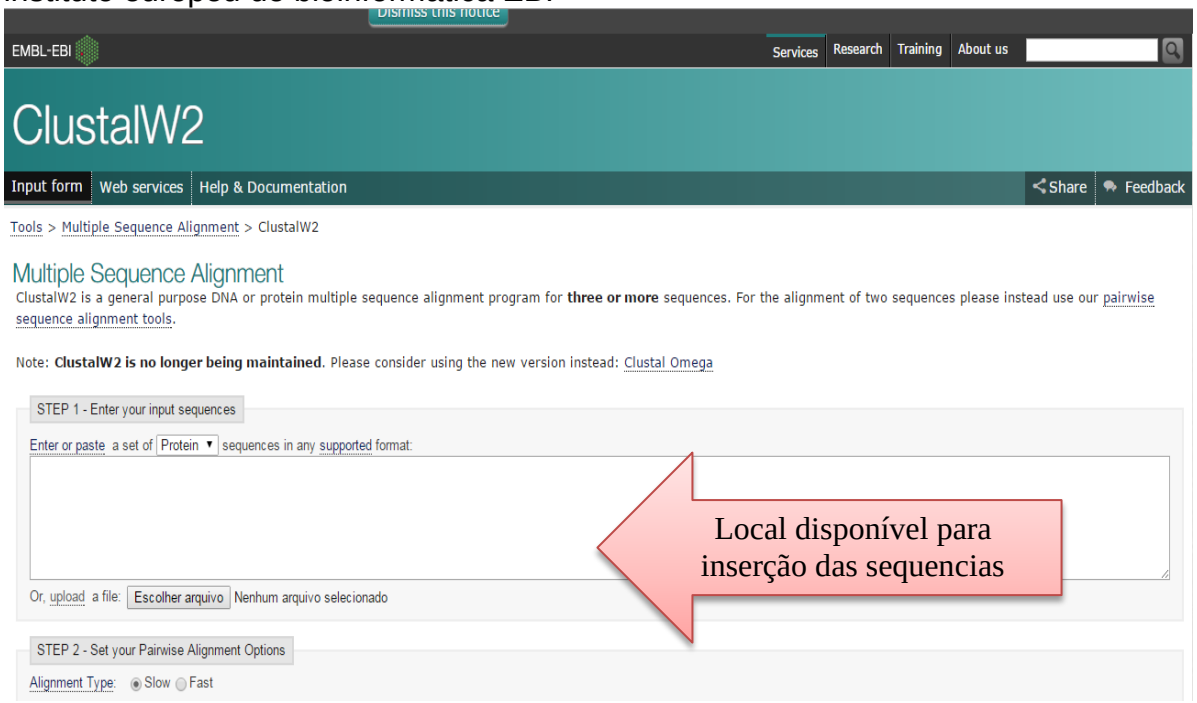
Entre os softwares relacionados na Tabela 1 foram escolhidos o CLUSTALW, MAFFT e T-COFFEE para serem realizadas as análises, por serem citados frequentemente por vários autores da área de pesquisa. Essas ferramentas são citadas e reconhecidas por utilizarem algoritmos considerados rápidos e com níveis de precisão significativos.

As três ferramentas escolhidas funcionam em ambiente web, evitando assim a instalação dos softwares na estação de trabalho. Como nem todos os softwares estão disponíveis na mesma plataforma, ou seja, uns funcionam no sistema operacional Windows e outros apenas no sistema operacional Linux, optou-se pelo ambiente web que não necessita de um sistema operacional específico, apenas de um navegador.

O conjunto de sequencias encontrados no NCBI e EBI foram submetidas à análise através das três ferramentas mencionadas, de modo a permitir a sua avaliação quanto à aplicabilidade, facilidade de estudo e viabilidade na análise de cada uma delas.

### 4.1.1 CLUSTALW

Figura 16 – Página inicial da ferramenta CLUSTALW disponível no servidor web do instituto europeu de bioinformática EBI




Fonte: European Molecular Biology Laboratory (2014b)

OBS: Dados trabalhados pelo autor

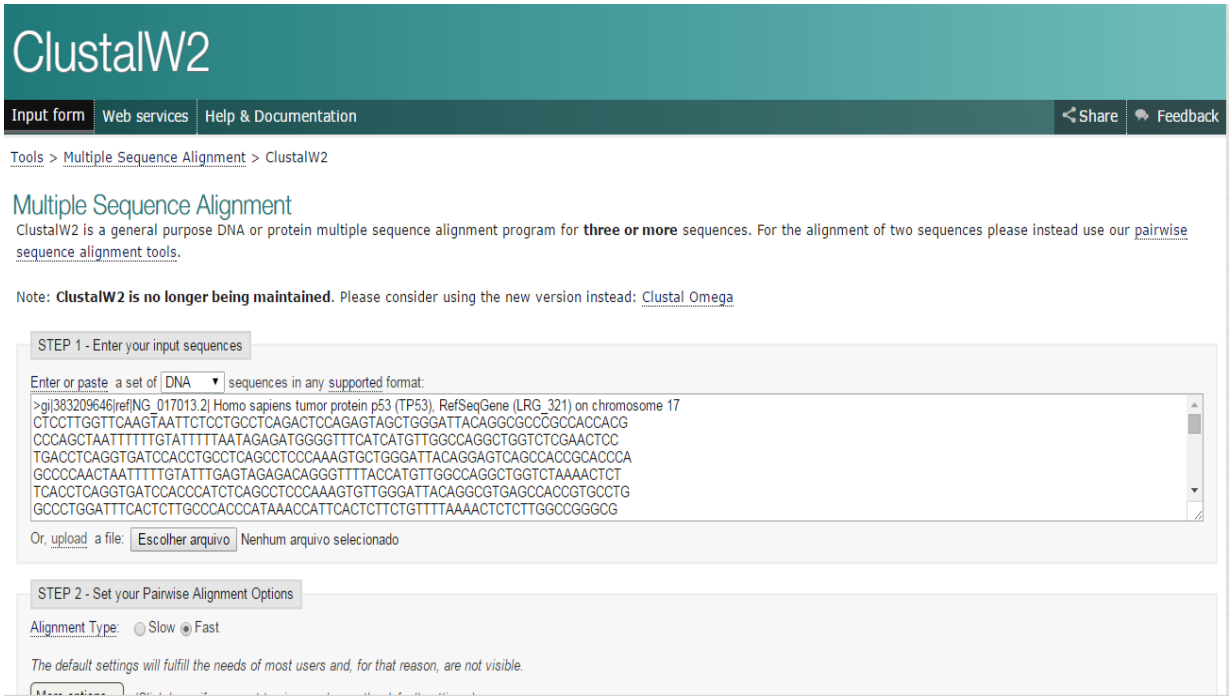
Para realizar a análise no CLUSTALW foi selecionada a sequencia original do gene p53 no banco de dados do NCBI e inserida no espaço indicado para as sequencias conforme apresentado na Figura 17.

A opção enviar o arquivo no formato texto com as sequencias foi ignorada, considerando que o número de sequencias a serem analisadas era razoavelmente pequeno. Para transpor as sequencias do arquivo texto, foi utilizado os comandos selecionar todas, copiar e colar no ambiente do CLUSTALW.

Abaixo da sequencia original do gene foram inseridas as sequências mutantes encontradas nos bancos de dados. As configurações da ferramenta não foram modificadas, exceto a opção tipo de alinhamento que foi escolhida a *fast*.

As sequencias encontradas no banco de dados do instituto europeu também foram submetidas à análise no CLUSTALW seguindo o mesmo procedimento descrito acima.

Figura 17 – Sequencias inseridas para análises na ferramenta CLUSTALW disponível no servidor web do instituto europeu de bioinformática EBI




Fonte: European Molecular Biology Laboratory (2014b)

OBS: Dados trabalhados pelo autor

Após a inserção das sequencias as mesmas foram submetidas à análise do CLUSTALW clicando no botão *submit.*

Em meados da década de 80, quando ainda a capacidade de processamento dos computares era muito inferior aos computadores de hoje, o processamento de análise múltiplo de sequencias era custoso e às vezes demandavam dias de processamento.

Para a quantidade de sequencias analisadas nesta pesquisa, o resultado foi obtido em menos de um minuto, pois as ferramentas utilizadas para alinhamento múltiplo de sequencias usam métodos heurísticos e não de otimização global. Algo que também deve ser levado em consideração na rapidez do processamento é o fato das ferramentas utilizadas nesta pesquisa estarem sendo executadas em computadores servidores, o que garante que são máquinas com poder de processamento maior do que um computador pessoal utilizado por um usuário domestico.

Analisando o resultado do CLUSTALW foi possível identificar a região do gene onde ocorre as principais mutações relacionadas ao câncer de pele. A

Figura 18 exibe o inicio da comparação entre a sequência do gene original e a sequencia da região mutante encontradas no NCBI. Cada traço representa a inserção de *gap*, ou seja, a região do gene que está fora da área comparada.

Figura 18 – Exemplo da exibição de resultados com a inserção de *gaps* da ferramenta CLUSTALW disponível no servidor web do instituto europeu de bioinformática EBI


Fonte: European Molecular Biology Laboratory (2014b)

OBS: Dados trabalhados pelo autor

Na Figura 19 é possível identificar na sequencia do gene original a região onde estão localizadas as mutações encontradas no banco de dados do NCBI para os canceres cutâneos.

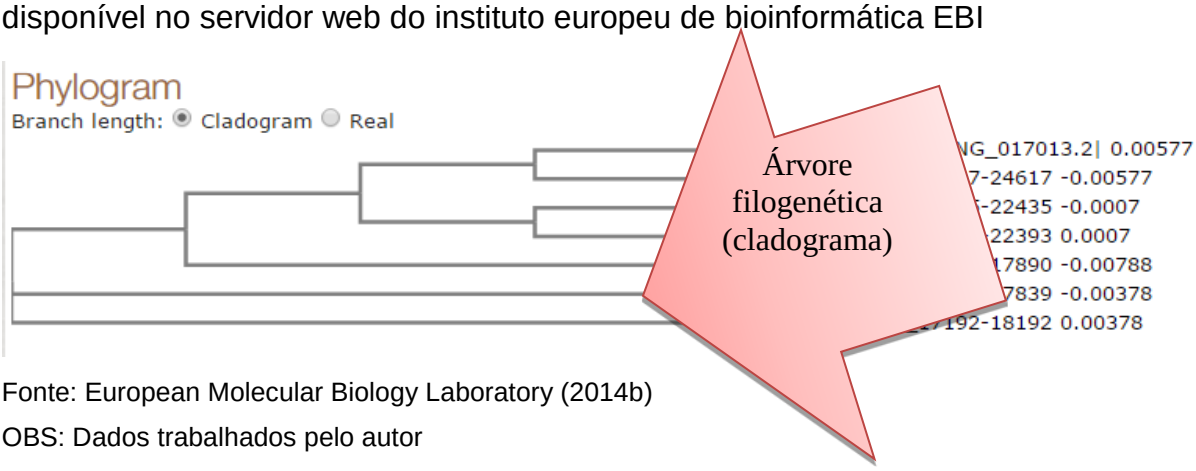
Figura 19 – Detalhes da análise entre a sequencia original e os mutantes disponibilizado pela ferramenta CLUSTALW armazenada no servidor web do instituto europeu de bioinformática EBI


Fonte: European Molecular Biology Laboratory (2014b)

OBS: Dados trabalhados pelo autor

Na Figura 20, indicado com uma seta, é possível visualizar o cladograma, também conhecido com a árvore filogenética que mostra as relação existentes entre as sequencias. Quanto mais distante da raiz, maior será a mutação.

Figura 20 – Exibição da arvore filogenética montada pela ferramenta CLUSTALW disponível no servidor web do instituto europeu de bioinformática EBI




Fonte: European Molecular Biology Laboratory (2014b)

OBS: Dados trabalhados pelo autor

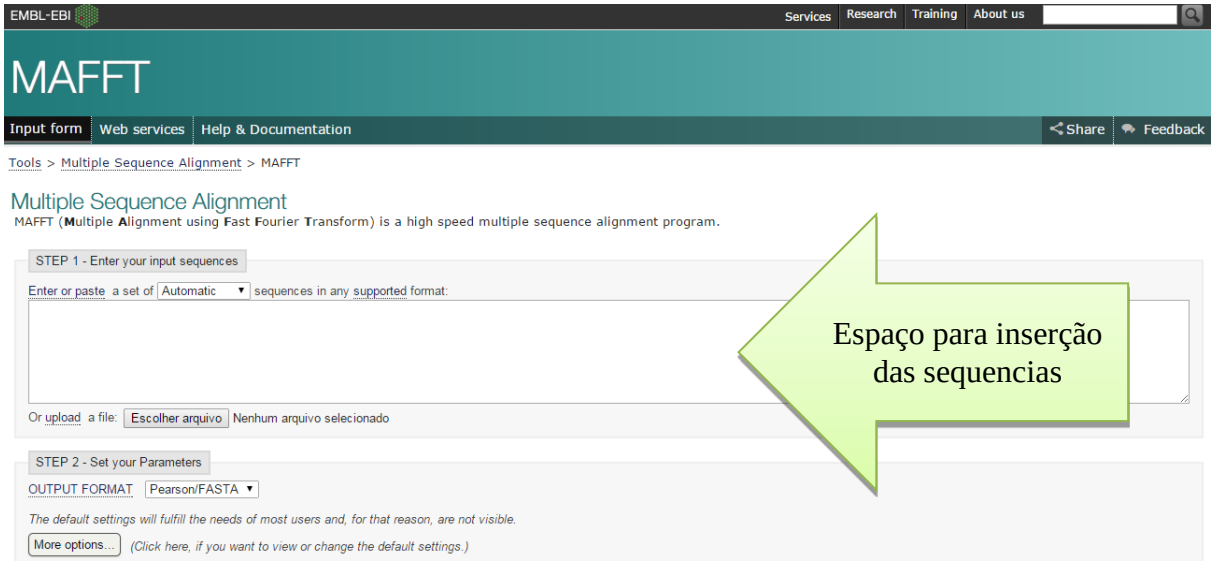
As ferramentas de análises múltiplas de sequencias estão disponíveis em vários servidores webs, portanto cada uma possui um ambiente distinto. Para o uso da ferramenta CLUSTALW, escolhemos a versão armazenada no servidor do instituto europeu de bioinformática – EBI. Também podemos encontrá-la em outros servidores com um layout de tela com algumas diferenças.

### 4.1.2 MAFFT

O MAFFT está localizado no servidor web do Instituto Europeu de Bioinformática (EBI), e foi a segunda ferramenta escolhida para realizar a análise com as mesmas sequencias que foram utilizadas na ferramenta CLUSTALW.

Os ambientes disponíveis na Web, das duas ferramentas de análises múltiplas de sequencias, são bastante semelhantes. Basicamente possuem um espaço para inserir as sequencias e um botão para *upload* do arquivo contendo as sequencias. Logo abaixo algumas opções de configurações e um botão *submit*.

Figura 21 – Página inicial da ferramenta MAFFT disponível no servidor web do instituto europeu de bioinformática EBI

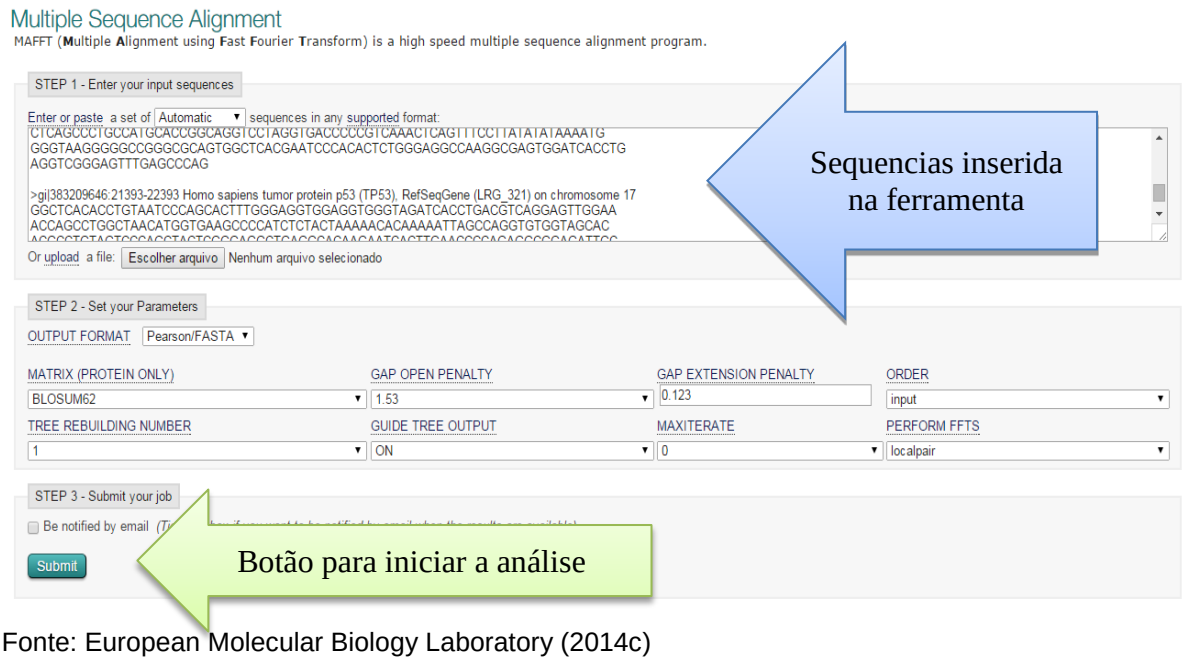



Fonte: European Molecular Biology Laboratory (2014c)

OBS: Dados trabalhados pelo autor

Foram inseridas no MAFFT tanto a sequencia original do gene p53 como as sequencias de mutantes localizadas no site do NCBI identificadas como de pacientes com câncer de pele. A Figura 22 exibe a tela com as sequencias já inseridas. Os parâmetros de configuração não foram modificados deixando para que o próprio software escolhesse a melhor opção.

Figura 22 – Sequencias inseridas para as análises na ferramenta MAFFT disponível no servidor web do instituto europeu de bioinformática EBI

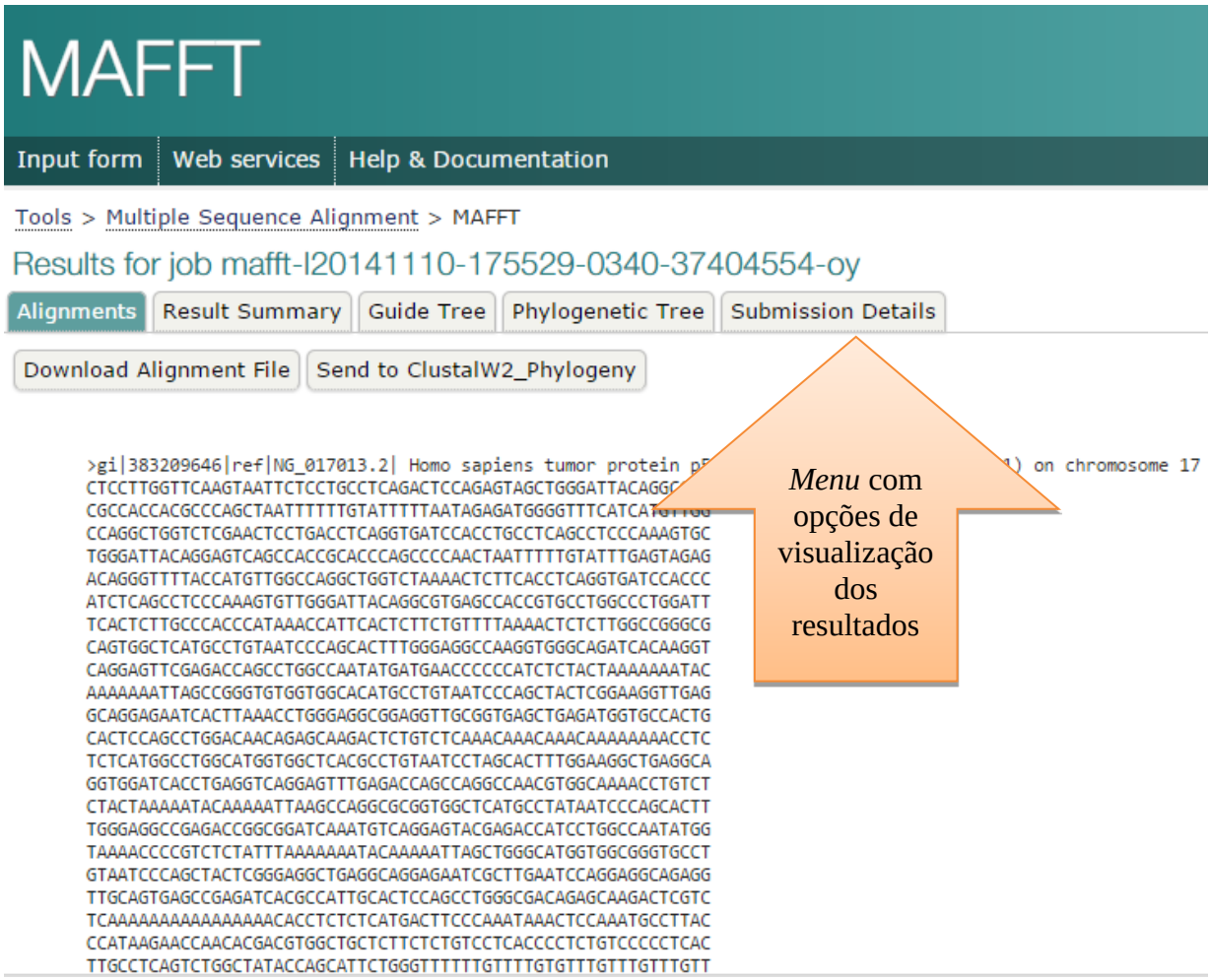



Fonte: European Molecular Biology Laboratory (2014c)

OBS: Dados trabalhados pelo autor

Conforme podemos identificar na Figura 23 o MAFFT apresentou o resultado com um layout diferente do CLUSTALW. Inicialmente ele apresentou a sequencia original sem mutação e logo abaixo apresentou as sequencias mutantes uma após a outra. A região das sequencias que estão fora da área de comparação, ou seja, onde uma sequencia é maior que a outra e representado por uma sequencia de caracteres identificados com um "-" conhecidos como *gaps*, conforme pode ser visto na Figura 24.

Figura 23 – Menu de resultados da ferramenta MAFFT disponível no servidor web do instituto europeu de bioinformática EBI




Fonte: European Molecular Biology Laboratory (2014c)

OBS: Dados trabalhados pelo autor

Figura 24 – Inserção de gaps no resultado da ferramenta MAFFT disponível no servidor web do instituo europeu de bioinformática EBI


Fonte: European Molecular Biology Laboratory (2014c)
OBS: Dados trabalhados pelo autor

Além da opção de exibição do cladograma - recurso disponível no CLUSTALW - o MAFFT exibe uma matriz de percentagem de identidade, conforme visto na Tabela 2, onde é possível verificar as diferenças entre as sequencias inseridas para análise. Os números na matriz indicam o grau de similaridade entre as sequencias. Para o numero 100, indica que as regiões das sequencias são idênticas. Já as colunas com '-nan' indica as regiões com *gaps*.

Tabela 2 - Matriz de percentagem de identidade gerada pela ferramenta MAFFT disponível no servidor web do instituto europeu de bioinformática EBI

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1: gi\|383209646\|ref\|NG_017013.2\| | 100.00 | 100.00 | 100.00 | 100.00 | 97.50 | 94.01 | 93.11 |
| 2: gi\|383209646_16839-17839 | 100.00 | 100.00 | 100.00 | 100.00 | 57.63 | 75.31 | 75.79 |
| 3: gi\|383209646_16890-17890 | 100.00 | 100.00 | 100.00 | 100.00 | 57.63 | 75.31 | 74.05 |
| 4: gi\|383209646_17192-18192 | 100.00 | 100.00 | 100.00 | 100.00 | 57.63 | -nan | -nan |
| 5: gi\|383209646_23617-24617 | 97.50 | 57.63 | 57.63 | 57.63 | 100.00 | -nan | -nan |
| 6: gi\|383209646_21435-22435 | 94.01 | 75.31 | 75.31 | -nan | -nan | 100.00 | 100.00 |
| 7: gi\|383209646_21393-22393 | 93.11 | 75.79 | 74.05 | -nan | -nan | 100.00 | 100.00 |

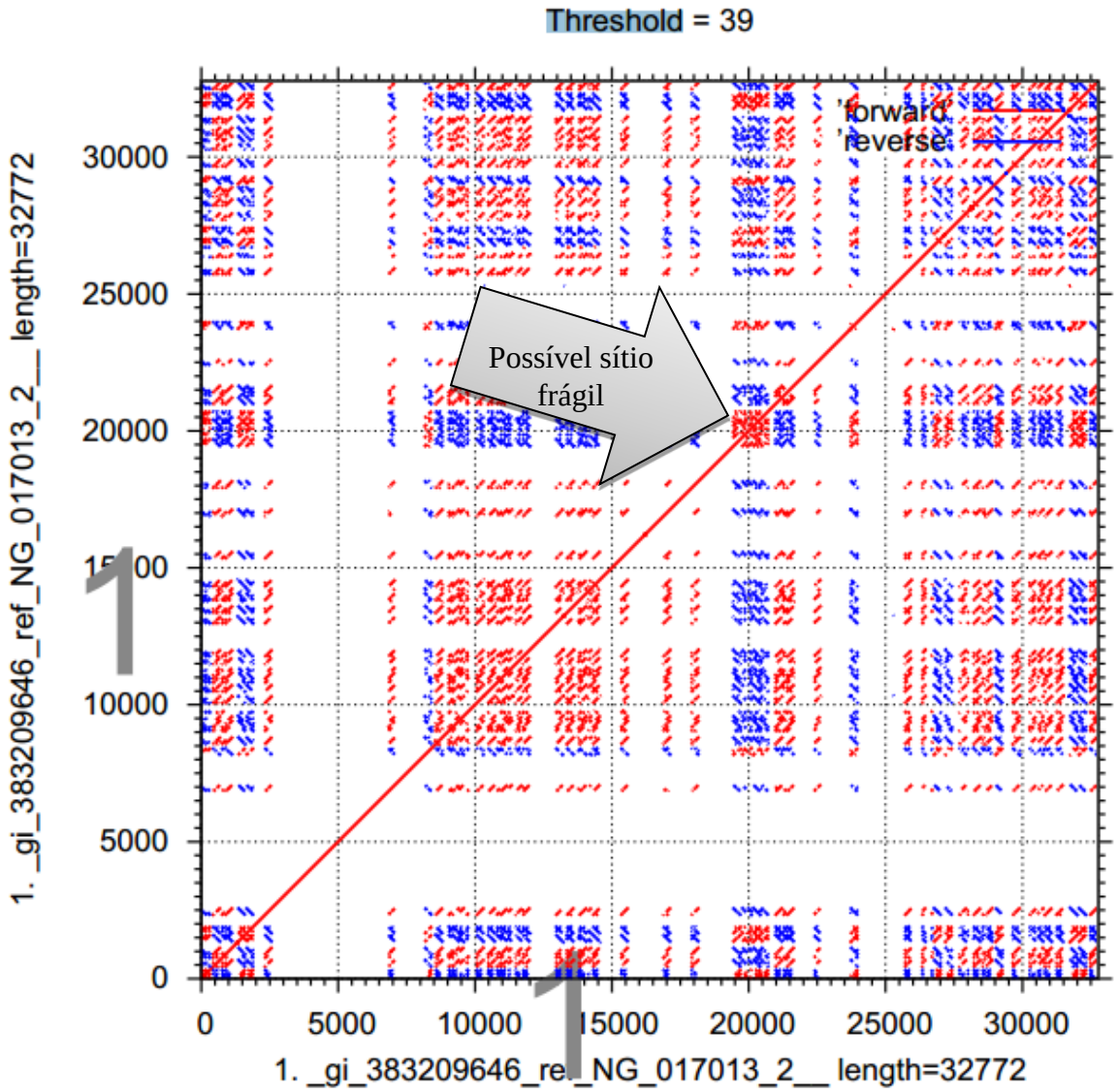
Fonte: European Molecular Biology Laboratory (2014c)

OBS: Dados trabalhados pelo autor

A fim de visualizar as mutações em um formato gráfico, foram realizadas análises das sequencias em questão na ferramenta MAFFT, que está armazenada no servidor web do Centro de Pesquisa em Biologia Computacional (CBRC) localizado no Japão. O software possui um layout de interface muito semelhante ao encontrado no servidor do EBI, com um recurso gráfico conforme visualizado na Figura 25. Nesta ilustração, regiões do gene p53 são identificadas contendo as mutações mais frequentemente depositadas no banco de dados do NCBI.

A Figura 25 exibe um retângulo preenchido com pontos coloridos. As dimensões do retângulo representa a maior sequencia, que neste caso, é a sequencia selvagem. Uma linha na diagonal cruza a imagem cortando a região com maior número de pontos que indica o local onde ocorrem as mutações, indicando um possível sítio frágil. Esta forma de apresentação de dados recebe o nome de *threshold* (ou limiar / borda de tolerância) das sequencias mutantes em relação a sequencia selvagem.

Figura 25 – *Threshold* gerado pela ferramenta MAFFT disponível no servidor web do Centro de Pesquisa em biologia Computacional CBRC com dados do NCBI




Fonte: European Molecular Biology Laboratory (2014c)

OBS: Dados trabalhados pelo autor

### 4.1.3 T-COFFEE

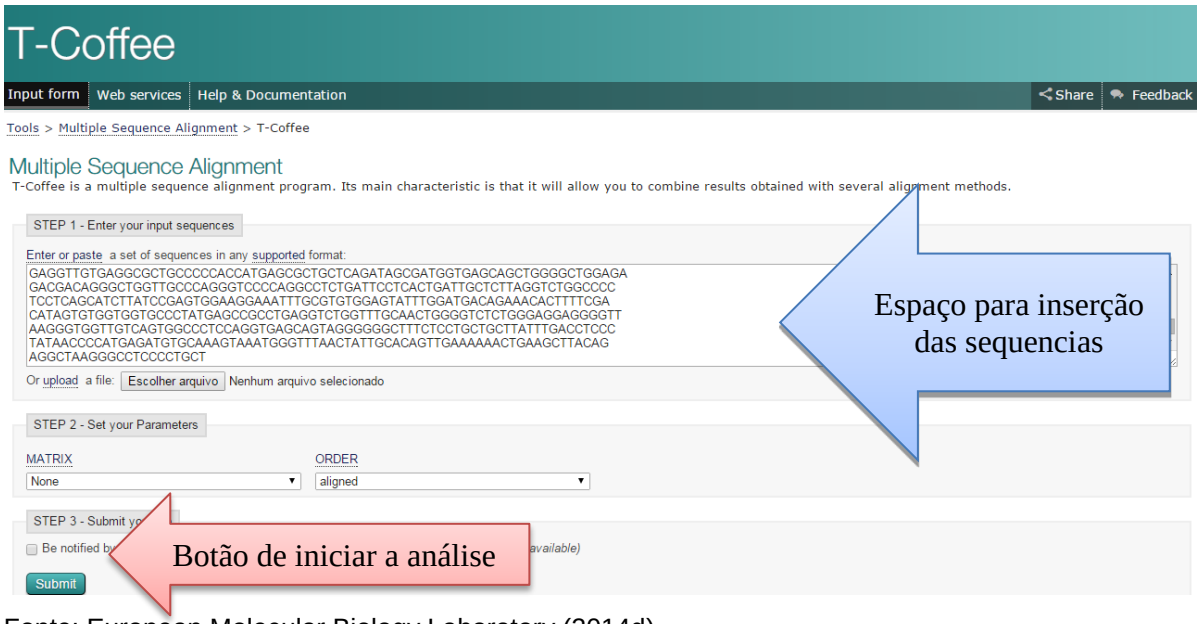
Foi também utilizada a ferramenta de análise múltipla de sequencia de nucleotídeos denominada T-COFFE, que se encontra disponível no EBI.

Seguindo o mesmo procedimento adotado para as ferramentas anteriores, foi inserido no T-COFFEE a sequencia original do gene p53 e as sequencias mutantes do mesmo gene encontradas no banco de dados do NCBI, conforme a Figura 26.

Foi adotado o mesmo procedimento das ferramentas anteriores

concernente aos parâmetros de configuração, não foram alterados, deixando que a o software escolhesse a melhor opção.

Figura 26 – Sequencias inseridas para análises na ferramenta T-COFFEEe disponível no servidor web do instituto europeu de bioinformática EBI




Fonte: European Molecular Biology Laboratory (2014d)

OBS: Dados trabalhados pelo autor

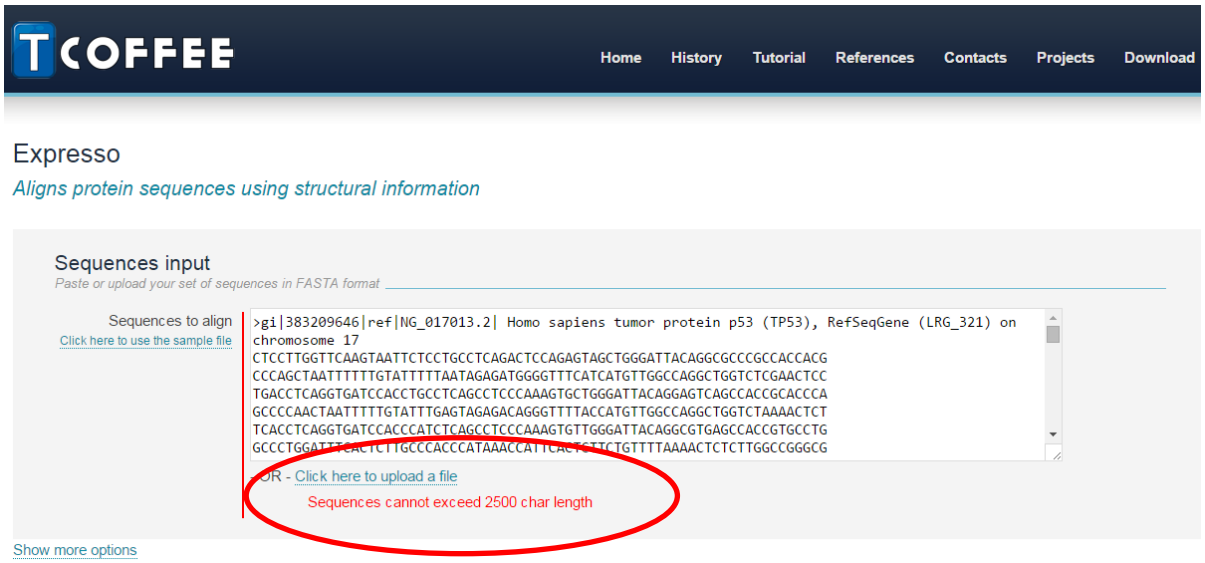
Ao submeter o alinhamento das sequencias na ferramenta T-COFFEE foi retornado um erro que não foi possível alinhar as sequencias inseridas conforme identificado na Figura 27.

Figura 27 – Mensagem de erro da ferramenta T-COFFEEe disponível no servidor web do instituto europeu de bioinformática EBI


Fonte: European Molecular Biology Laboratory (2014d)

Como a mensagem de erro não foi esclarecedora foi realizado a mesma análise no T-COFFEE armazenado no servidor web de domínio http://tcoffee.crg.cat/ que retornou claramente o motivo pelo qual não foi possível realizar a análise. As sequencias inseridas no T-COFFEE não podem ultrapassar o limite de 2500 caracteres conforme Figura 28.

Figura 28 – Mensagem de erro da ferramenta T-COFFEEe disponível no servidor web do centro de regulação genômica de Barcelona CRG

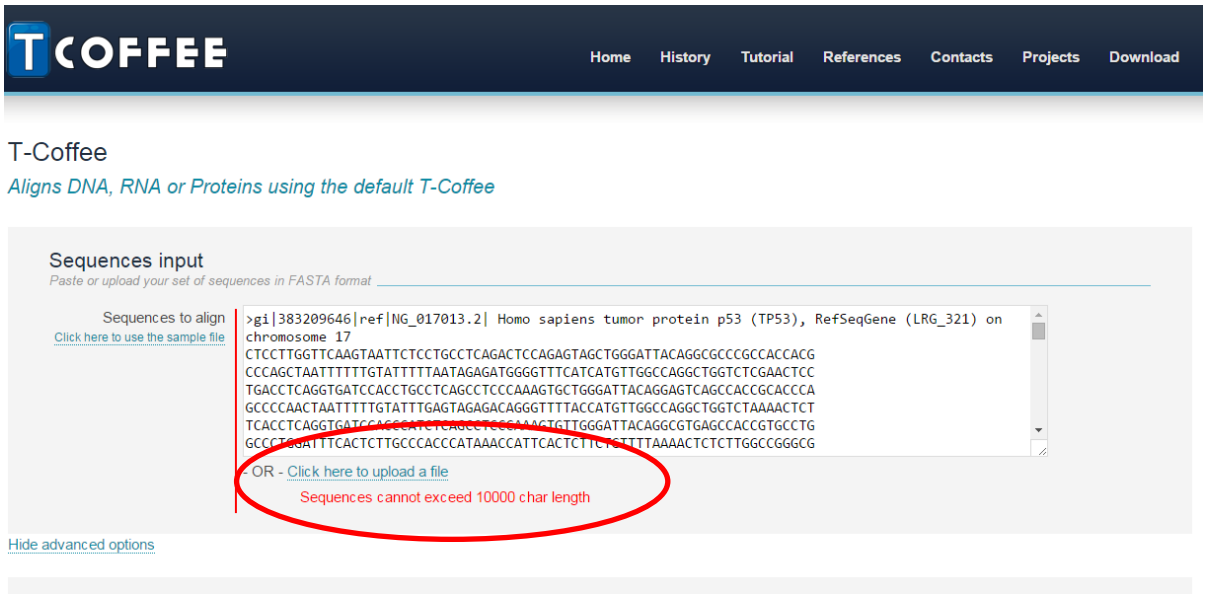



Fonte: European Molecular Biology Laboratory (2014d)

OBS: Dados trabalhados pelo autor

Também foi realizada a submissão das mesmas sequencias encontradas no NCBI e analisadas pelo CLUSTALW e MAFFT no T-COFFEE armazenado em outro servidor com domínio www.tcoffee.org. Mesmo possuindo uma capacidade maior para a quantidade de caracteres entrantes não foi possível realizar as análises já que o grupo de sequencias encontradas nos bancos de dados possui uma quantidade de caracteres superior a 39 mil para a sequência selvagem e os mutantes e a ferramenta está limitada apenas a 10 mil caracteres conforme podemos verificar na Figura 29.

Figura 29 – Mensagem de erro da ferramenta T-COFFEEe disponível no servidor web do laboratório de Notredame's




Fonte: European Molecular Biology Laboratory (2014d)

OBS: Dados trabalhados pelo autor

As ferramentas utilizadas, CLUSTALW, MAFFT e T-COFFEE estão disponíveis em mais de um servidor web, todos com as mesmas funcionalidades, mudando apenas o layout de apresentação da página inicial. Nas ferramentas de análise múltiplos de sequencias, apresentadas nesta pesquisa é possível, antes de clicar no botão *submit*, alterar algumas configurações ou deixar que o próprio software escolha a melhor opção para a análise.

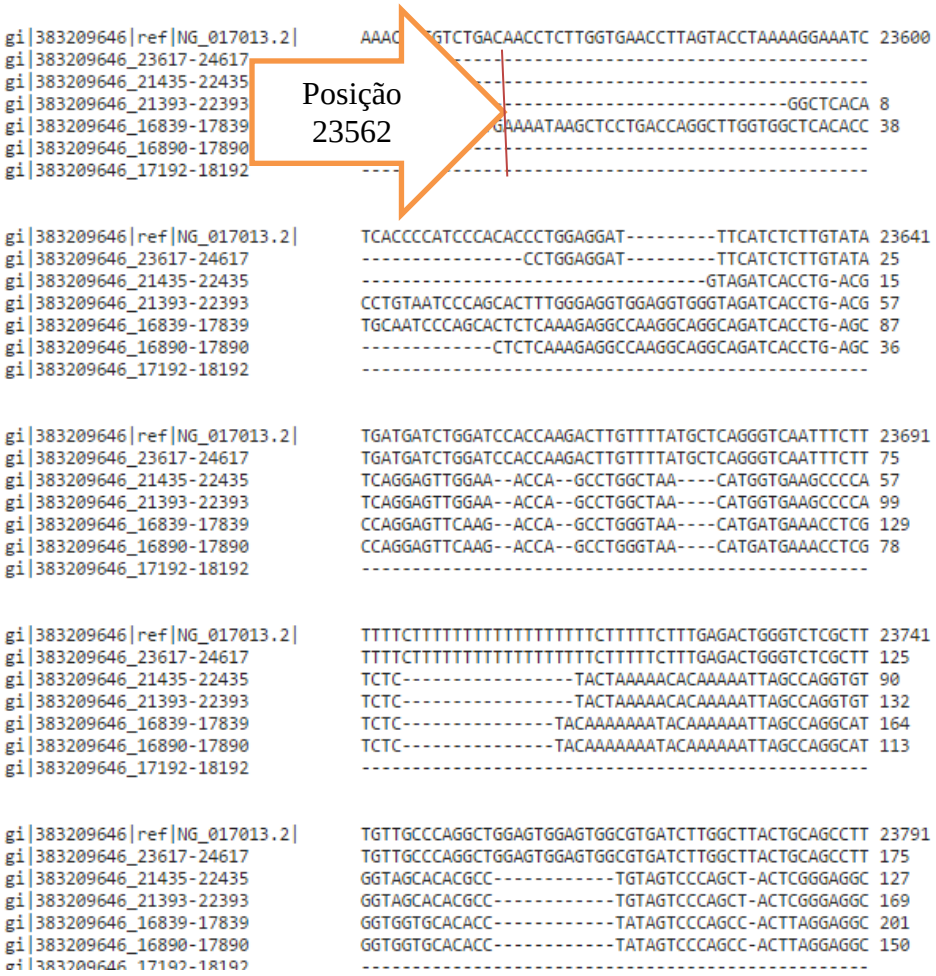
## 4.2 Resultados do CLUSTALW e MAFFT

As sequencias mutantes encontradas no NCBI e EBI foram submetidas separadamente ao CLUSTALW e MAFFT. Também foi criado um arquivo do tipo texto contendo uma sequencia original do gene p53 e todas as sequencias mutantes encontradas nos bancos de dados do NCBI e EBI.

### 4.2.1 Sequencias oriundas da base de dados do NCBI

Observando os resultados das análises realizadas pelas ferramentas CLUSTALW e MAFFT para as sequencias obtidas no NCBI foi possível identificar claramente que as mutações ocorrem entre as posições 23562 e 24735 do gene p53 conforme visto, respectivamente, nas Figura 30 e Figura 31.

Figura 30 – Posição inicial da mutação gerada pela ferramenta CLUSTALW disponível no EBI com os dados encontrados na base de dados americana NCBI




```
gi|383209646|ref|NG_017013.2|      AAAC TGTCTGACAACCTCTTGGTGAACCTTAGTACCTAAAAGGAAATC 23600
gi|383209646_23617-24617           ----------------------------------------          
gi|383209646_21435-22435           ----------------------------------------          
gi|383209646_21393-22393           ----------------------------GGCTCACA 8
gi|383209646_16839-17839           GAAAATAAGCTCCTGACCAGGCTTGGTGGCTCACACC 38
gi|383209646_16890-17890           ----------------------------------------          
gi|383209646_17192-18192           ----------------------------------------          

gi|383209646|ref|NG_017013.2|      TCACCCCATCCCACACCCTGGAGGAT---------TTCATCTCTTGTATA 23641
gi|383209646_23617-24617           ----------------CCTGGAGGAT---------TTCATCTCTTGTATA 25
gi|383209646_21435-22435           ----------------------------------GTAGATCACCTG-ACG 15
gi|383209646_21393-22393           CCTGTAATCCCAGCACTTTGGGAGGTGGAGGTGGGTAGATCACCTG-ACG 57
gi|383209646_16839-17839           TGCAATCCCAGCACTCTCAAAGAGGCCAAGGCAGGCAGATCACCTG-AGC 87
gi|383209646_16890-17890           -------------CTCTCAAAGAGGCCAAGGCAGGCAGATCACCTG-AGC 36
gi|383209646_17192-18192           ----------------------------------------------------

gi|383209646|ref|NG_017013.2|      TGATGATCTGGATCCACCAAGACTTGTTTTATGCTCAGGGTCAATTTCTT 23691
gi|383209646_23617-24617           TGATGATCTGGATCCACCAAGACTTGTTTTATGCTCAGGGTCAATTTCTT 75
gi|383209646_21435-22435           TCAGGAGTTGGAA--ACCA--GCCTGGCTAA----CATGGTGAAGCCCCA 57
gi|383209646_21393-22393           TCAGGAGTTGGAA--ACCA--GCCTGGCTAA----CATGGTGAAGCCCCA 99
gi|383209646_16839-17839           CCAGGAGTTCAAG--ACCA--GCCTGGGTAA----CATGATGAAACCTCG 129
gi|383209646_16890-17890           CCAGGAGTTCAAG--ACCA--GCCTGGGTAA----CATGATGAAACCTCG 78
gi|383209646_17192-18192           ----------------------------------------------------

gi|383209646|ref|NG_017013.2|      TTTTCTTTTTTTTTTTTTTTTTTCTTTTTCTTTGAGACTGGGTCTCGCTT 23741
gi|383209646_23617-24617           TTTTCTTTTTTTTTTTTTTTTTTCTTTTTCTTTGAGACTGGGTCTCGCTT 125
gi|383209646_21435-22435           TCTC-----------------TACTAAAAACACAAAAATTAGCCAGGTGT 90
gi|383209646_21393-22393           TCTC-----------------TACTAAAAACACAAAAATTAGCCAGGTGT 132
gi|383209646_16839-17839           TCTC---------------TACAAAAAAATACAAAAAATTAGCCAGGCAT 164
gi|383209646_16890-17890           TCTC---------------TACAAAAAAATACAAAAAATTAGCCAGGCAT 113
gi|383209646_17192-18192           ----------------------------------------------------

gi|383209646|ref|NG_017013.2|      TGTTGCCCAGGCTGGAGTGGAGTGGCGTGATCTTGGCTTACTGCAGCCTT 23791
gi|383209646_23617-24617           TGTTGCCCAGGCTGGAGTGGAGTGGCGTGATCTTGGCTTACTGCAGCCTT 175
gi|383209646_21435-22435           GGTAGCACACGCC------------TGTAGTCCCAGCT-ACTCGGGAGGC 127
gi|383209646_21393-22393           GGTAGCACACGCC------------TGTAGTCCCAGCT-ACTCGGGAGGC 169
gi|383209646_16839-17839           GGTGGTGCACACC------------TATAGTCCCAGCC-ACTTAGGAGGC 201
gi|383209646_16890-17890           GGTGGTGCACACC------------TATAGTCCCAGCC-ACTTAGGAGGC 150
gi|383209646_17192-18192           ----------------------------------------------------
```

Fonte: European Molecular Biology Laboratory (2014b)

OBS: Dados trabalhados pelo autor

Figura 31 - Posição final da mutação gerada pela ferramenta CLUSTALW disponível no EBI com os dados encontrados na base de dados americana NCBI


Fonte: European Molecular Biology Laboratory (2014b)

OBS: Dados trabalhados pelo autor

Não foi possível estabelecer uma frequência de mutações ocorridas nos nucleotídeos entre as seis mutações encontradas no banco de dados do NCBI conforme identificados na Tabela 3. As substituições das bases não ocorrem em uma frequência que seja possível identificá-las.

A Tabela 3 foi construída com a finalidade de expor o cabeçalho de cada mutante encontrado no banco de dados do NCBI. Na primeira linha é apresentado o nome do mutante, na segunda o fenótipo, na terceira a localização e na última linha um código identificador.

Tabela 3- Cabeçalho dos Mutantes encontrados na base de dados americana NCBI

| |
|---|
| 1. Name: NM_000546.5(TP53):c.400T>C (p.Phe134Leu)<br>Condition(s): Malignant melanoma<br>Location (GRCh38): 7675212<br>ID: 80710 |
| 2. Name: NM_000546.5(TP53):c.672G>A (p.Glu224=)<br>Condition(s): Malignant melanoma<br>Location (GRCh38): 7674859<br>ID: 80709 |
| 3. Name: NM_000546.5(TP53):c.1093C>T (p.His365Tyr)<br>Condition(s): Malignant melanoma, Neoplastic Syndromes, Hereditary<br>Location (GRCh38): 7670616<br>ID: 80708 |
| 4. Name: NM_000546.5(TP53):c.1051A>G (p.Lys351Glu)<br>Condition(s): Malignant melanoma, Neoplastic Syndromes, Hereditary<br>Location (GRCh38): 7670658<br>ID: 72340 |
| 5. Name: NM_000546.5(TP53):c.*1175A>C<br>Condition(s): Basal cell carcinoma, susceptibility to, 7<br>Location (GRCh38): 7668434<br>ID: 35555 |
| 6. Name: NM_000546.5(TP53):c.451C>A (p.Pro151Thr)<br>Condition(s): Breast cancer, somatic, Breast adenocarcinoma, Neoplastic Syndromes, Hereditary<br>Location (GRCh38): 7675161<br>ID: 12369 |

Fonte: Estados Unidos (2014)

OBS: Dados trabalhados pelo autor

Na Tabela 4 está apresentado o endereço e nome de cada laboratório ou instituição de onde proveio a sequencia mutante depositada no banco de dados do NCBI. A coluna ID da Tabela 4 refere-se o gene mutante numerado sequencialmente na Tabela 3. Foi possível constatar que a mesma mutação foi detectada em laboratórios diferentes.

Tabela 4 – Identificação dos Laboratórios de origem dos mutantes encontrados na base de dados americana NCBI

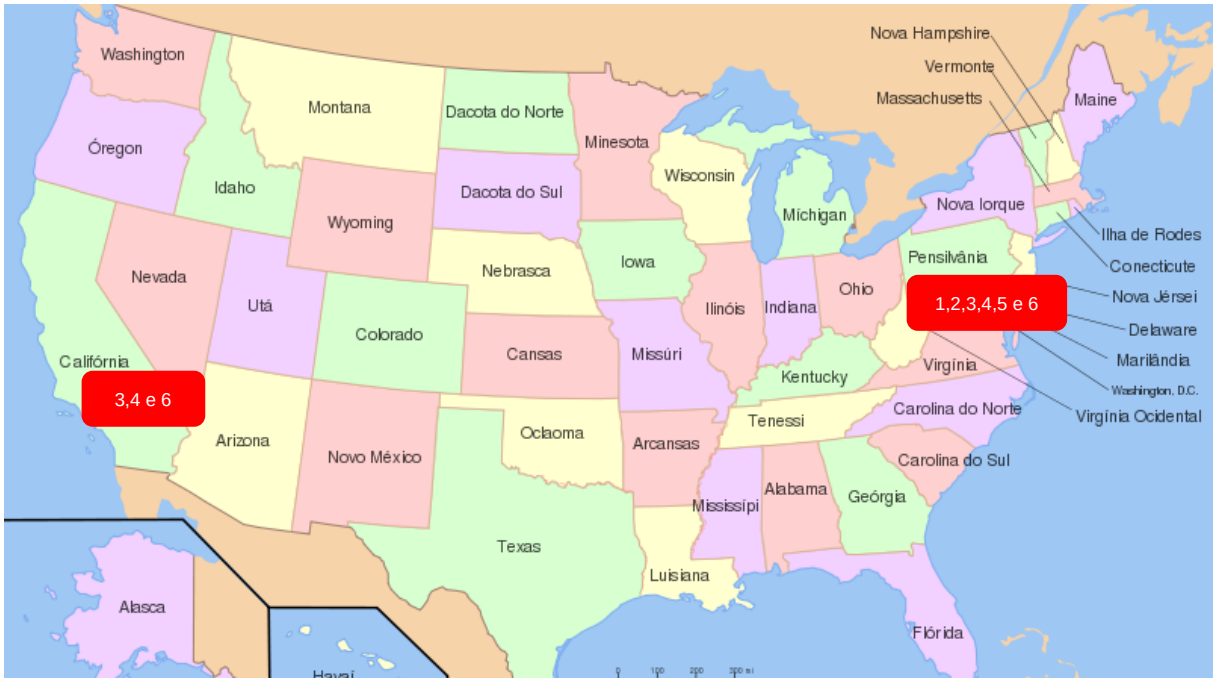
| Mutante (NCBI) | ID | Laboratório envolvido com o sequenciamento |
|---|---|---|
| 80710<br>80709<br>80708<br>72340 | 1<br>2<br>3<br>4 | Samuels Laboratory (Organization ID: 500168)<br>NHGRI/NIH<br>50 South Drive<br>MSC 8000, Building 50, Rm 5140<br>Bethesda Maryland<br>United States - 20892-8000<br>(BOLLAG ET AL. 2010)<br>(BUON, 2013) |
| 80708<br>72340<br>12369 | 3<br>4<br>6 | Ambry Genetics (Organization ID: 61756)<br>Ambry Genetics<br>15 Argonaut<br>Aliso Viejo<br>California<br>United States – 92656<br>(BOLLAG ET AL. 2010)<br>(BUON, 2013) |
| 35555<br>12369 | 5<br>6 | OMIM (Organization ID: 3)<br>Johns Hopkins University<br>Baltimore<br>Maryland<br>United States<br>(PARKIN, 2004)<br>(BUON, 2013) |

Fonte: Estados Unidos (2014)

OBS: Dados trabalhados pelo autor

Na Figura 32 foi possível identificar geograficamente o laboratório onde foi coletado o tecido do paciente diagnosticado com câncer de pele. Não há informações pessoais sobre nenhum paciente, apenas está disponível informações sobre o sequenciamento do material coletado.

Figura 32 - Indicação geográfica da localização do laboratório de origem do mutante encontrado na base da dados americana NCBI



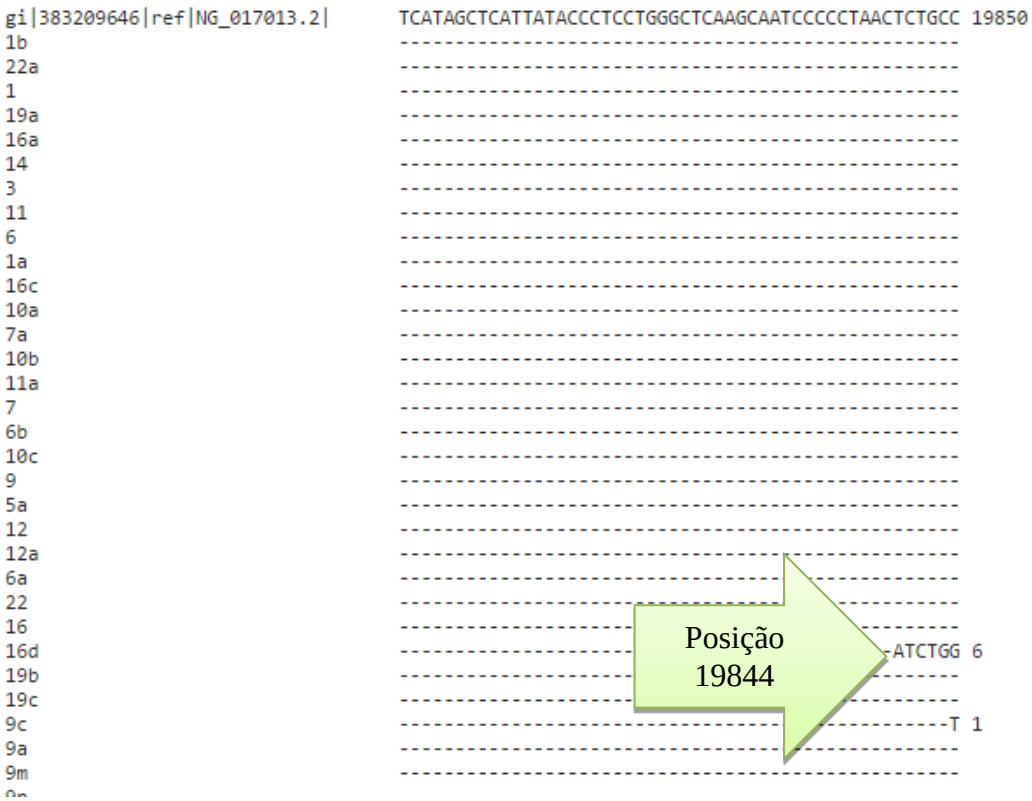

Fonte: MAPA... (2014)

OBS: Dados trabalhados pelo autor

### 4.2.2 Sequencias oriundas da base de dados do EBI

Observando os resultados das análises realizadas pelas ferramentas CLUSTALW E MAFFT para as sequencias obtidas no EBI foi possível identificar claramente que as mutações ocorrem entre as posições 19844 e 21786 do gene p53 conforme visto, respectivamente, na Figura 33 e Figura 34.

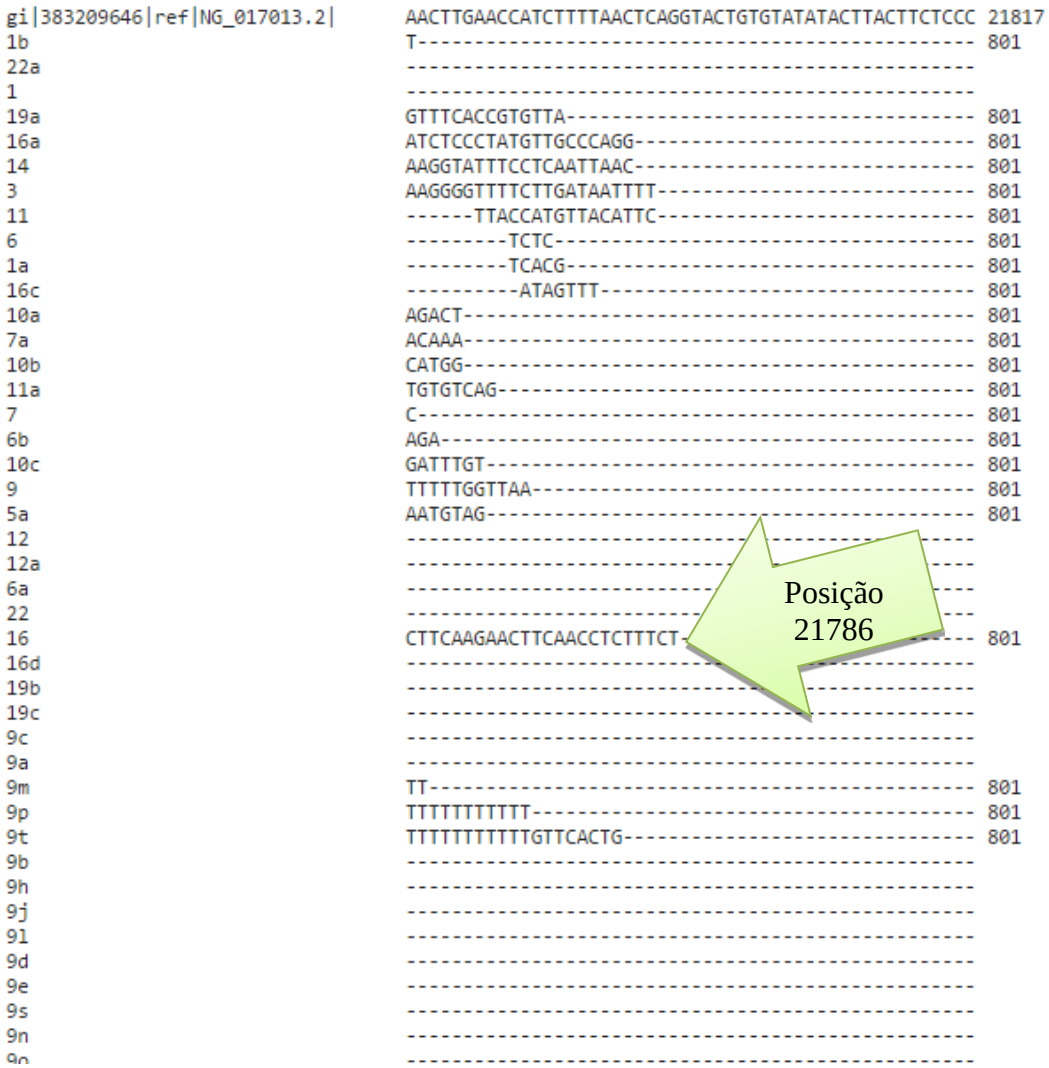
Figura 33 - Posição inicial da mutação gerada pela ferramenta CLUSTALW disponível no EBI com os dados encontrados na base de dados europeia EBI




Fonte: European Molecular Biology Laboratory (2014b)

OBS: Dados trabalhados pelo autor

Figura 34 - Posição final da mutação gerada pela ferramenta CLUSTALW disponível no EBI com os dados encontrados na base de dados europeia EBI

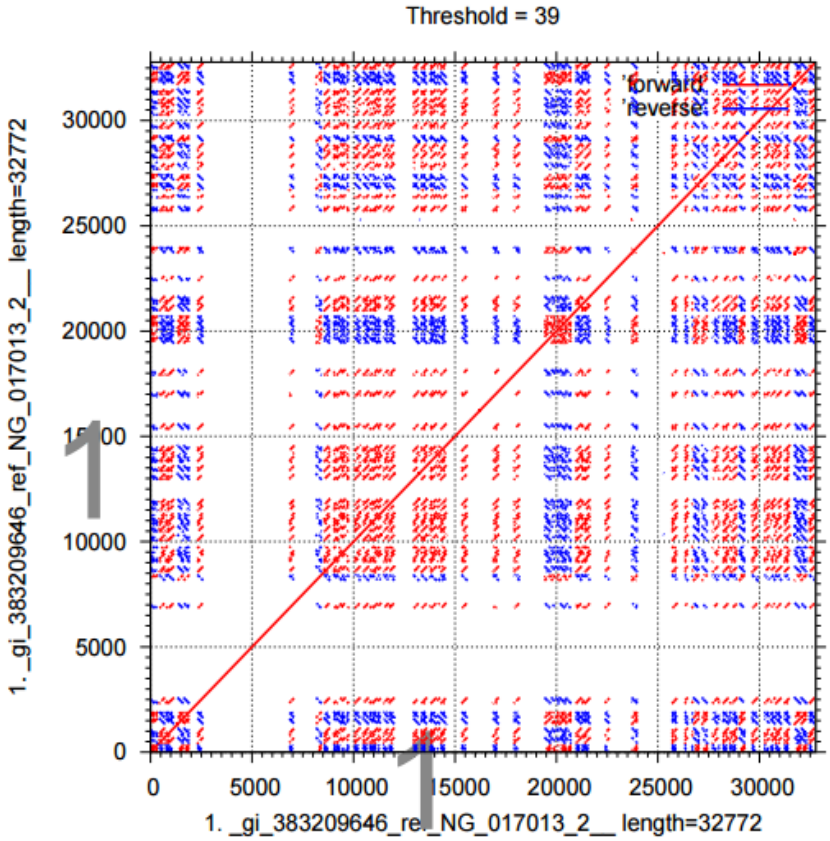



Fonte: European Molecular Biology Laboratory (2014b)

OBS: Dados trabalhados pelo autor

A disposição dos dados no banco de dados do EBI não estão no mesmo formato encontrado no NCBI. A Tabela 5 foi montada para facilitar a visualização dos códigos de cada mutante e a identificação da transformação ocorrida. Identificou-se que não houve mutação de deleção de nucleotídeos, apenas substituições. As mais frequentes foram as trocas de citosina por timina e guanina por adenina, ambas com as mesmas quantidades, conforme visto na Tabela 6.

Figura 35 - Threshold gerado pela ferramenta MAFFT disponível no servidor web do Centro de Pesquisa em biologia Computacional CBRC com dados do EBI




Fonte: Japão (2014)

Tabela 5 – Nomes dos mutantes encontrados na base de dados europeia EBI relacionados com câncer de pele

| | |
|---|---|
| rs12203592 - C/T | rs45430 - C/T |
| rs1805007 - C/G/T | rs7023329 - A/G |
| rs12202284 - C/A | rs401681 - C/T |
| rs8015138 - A/C | rs3219090 - T/C |
| rs113488022 - A/G/T/C | rs228437 - C/T |
| rs137853080 - T/G | rs35390 - C/A |
| rs137853081 - G/C | rs1722784 - A/G |
| rs121913323 - C/T | rs6001027 - A/G |
| rs121913315 - G/A/T | rs36204594 - G/A |
| rs202187871 - C/T | rs11552822 - C/A |
| rs149617956 - G/A | rs11547328 - G/A |
| rs16891982 - C/G | rs104894098 - A/T |
| rs121909233 - G/A | rs113798404 - C/A/G |
| rs121909232 - C/A | rs104894094 - C/A |
| rs121909234 - G/A | rs137854599 - C/T |
| rs258322 - A/G | rs137854597 - C/T |
| rs4785763 - A/C | rs121913388 - G/A |
| rs910873 - G/A | rs104894099 - A/C |
| rs1393350 - G/A | rs104894109 - C/A |
| rs16953002 - G/A | rs104894095 - C/G/T |
| rs17119461 - T/C | rs104894097 - C/G |
| rs7412746 - C/T | rs202187871 - C/T |
| rs13016963 - A/G | rs104894340 - C/T |
| rs2284063 - A/G | rs1805006 - C/A |
| rs1801516 - G/A | rs861539 - G/A |

Fonte: European Molecular Biology Laboratory (2014a)

OBS: Dados trabalhados pelo autor

Para analisar a frequência das mutações foi considerada apenas a substituição de um nucleotídeo por outro. As mutações diferentes não foram consideradas.

Tabela 6 – Frequência das substituições de nucleotídeos para os mutantes encontrados na base de dados do instituto europeu EBI

| | |
|---|---|
| T/G | 1 |
| G/C | 1 |
| A/T | 1 |
| T/C | 2 |
| C/G | 2 |
| A/C | 3 |
| A/G | 6 |
| C/A | 6 |
| C/T | 11 |
| G/A | 11 |

Fonte: Autor

### 4.2.3 Sequencias oriundas da base de dados do DDBJ

Na base de dados do DDBJ foram localizadas quatro mil e setenta e cinco sequencias mutantes para o gene p53. Infelizmente, não foi possível filtrar ou selecionar apenas as sequencias extraída de pacientes com câncer de pele, relacionadas ao gene em questão. Este fato impossibilitou a identificação de mutantes específicos do gene p53 e que estavam relacionados unicamente à patologia em questão. Este fato inviabilizou a realização da análise comparativa das sequências de DNA obtidas, tal como foi realizado a partir de sequências oriundas dos outros dois bancos de dados genômicos. Deste modo, infelizmente, não foi possível obter resultados a partir da aplicação das ferramentas CLUSTALW e MAFFT para as sequencias do gene p53 oriundas do banco de dados do Japão (DDBJ).

# 5 DISCUSSÃO

Os resultados obtidos na presente pesquisa demonstraram a viabilidade e exequibilidade das metodologias de análise propostas. As pesquisas nos bancos de dados NCBI e EBI permitiram a obtenção das sequências nucleotídicas do gene p53 (sequência original), bem como de mutantes deste gene, originados de pacientes com o diagnóstico clínico comprovado para câncer de pele.

Foi realizada a análise comparativa das sequências original e mutantes através da ferramenta BLAST, uma vez que este método permite a obtenção rápida de sequências mutantes para o gene p53, ou outras sequências com alto grau de similaridade (ALTSCHUL; MADDEN; SCHÄFFER, 1997). Entretanto, foi notado que tal procedimento seria inadequado para o propósito da pesquisa, uma vez que as sequências variantes não estariam, necessariamente, relacionadas com a doença em questão – câncer de pele.

Deste modo, optou-se pela utilização do recurso de identificação de variantes das sequências, disponibilizado pelo próprio banco de dados (NCBI e EBI). Esta estratégia mostrou-se mais eficiente e permitiu a identificação dos acessos de sequências mutantes contendo a indicação do fenótipo dos pacientes dos quais as mesmas foram isoladas. Deste modo, foram selecionados unicamente aqueles que possuíam relação direta com os casos comprovados de câncer de pele.

Todas as sequencias mutantes encontradas no banco de dados do NCBI estão relacionadas diretamente aos cânceres de pele, não levando em consideração os variados tipos de cânceres encontrados na literatura.

De acordo com Santos (2004) a observação dos dados analisados pelas ferramentas CLUSTALW e MAFFT pode ser útil para identificar possíveis sítios frágeis em qualquer gene. Foi possível identificar que as mutações ocorridas no gene p53, provenientes de material genético sequenciado de pacientes oriundos dos Estados Unidos, estão localizadas em um pequena região gênica. Considerando que a sequencia original de nucleotídeos do gene p53 é composta por 32 mil bases os mutantes foram identificados entre as posições 23562 e 24735.

A utilização das ferramentas CLUSTALW e MAFFT permitiu a identificação dos sítios de mutação em cada variante identificado, bem como a classificação das variações mais comuns. Este resultado permite constatar que a análise individual das sequências é uma poderosa ferramenta para a identificação

de possíveis sítios frágeis do genoma mais propensas a sofrer alterações e que, por sua vez, estariam correlacionadas com determinadas patologias.

Esta observação permite supor que a região em questão seja considerada um possível “sítio frágil” para a referida doença (câncer de pele), e que possa existir uma correlação entre variações ocorridas nesta, relativamente, pequena região do gene p53 e a consequente manifestação da doença nos pacientes de onde as mesmas foram isoladas.

As sequencias mutantes encontradas no banco de dados do NCBI e EBI se caracterizaram como mutantes em função da troca de um nucleotídeo por outro (substituição), ou mesmo deleções ou inserções de nucleotídeos. Comparando os resultados da frequência das mutações do banco de dados do NCBI e EBI foi possível identificar que os mutantes do gene p53 possuem mutações bem semelhantes. Em ambos, é mais comum a mutação do tipo transição.

Apesar no número de sequências mutantes ser considerado relativamente pequeno, foi notado que, para as sequências encontradas no banco de dados do NCBI a frequência mais comum de variação foi do tipo transição, ou seja, a troca entre purinas e purinas ou pirimidinas e pirimidinas. A transição do tipo purinas é a substituição de adenina por guanina ou guanina por adenina, e as transições do tipo pirimidinas é a substituição de citosina por timina ou timina por citosina, fato este que é respaldado pelos relatos na literatura (WATSON et al., 2006).

Dentre as seis sequencias mutantes encontradas no banco de dados do NCBI 4 foram do tipo transição e 2 do tipo transversão. Com esse número é possível afirmar que cerca de 66% das mutações são do tipo transição e 34% do tipo transversão. As substituições do tipo transversão se caracterizam pela substituição de purinas por pirimidinas ou pirimidinas por purinas.

No banco de dados do EBI foi possível analisar um número maior se sequencias mutantes. De forma análoga ao caso anterior, verificou-se que das 50 sequencias mutantes, 44 se caracterizaram como substituições, dentre estas, 30 foram do tipo transição e 14 do tipo transversão, ou seja, 68% transição e 31% transversão, semelhante ao banco de dados NCBI e em concordância com os relatos na literatura (WATSON et al., 2006).

As sequencias mutantes encontradas nesta pesquisa foram inseridas no banco de dados do NCBI por três laboratórios dos Estados Unidos. Dois desses

laboratórios estão localizados em Maryland (BOLLAG et al. 2010) e (BUON, 2013) e o outro na Califórnia (BUON, 2013).

Observa-se na Figura 32 que a mesma mutação foi encontrada em laboratórios diferentes. Considerando o ID da tabela como o identificados da sequencia mutante, foi possível verificar que a mutação 80710 e 80709 foi encontrada apenas em Bethesda Maryland, já a mutação 80708 e 72340 foi encontrada em Bethesda Maryland e também na Califórnia. A mutação 35555 foi identificada apenas em Baltimore Maryland e a mutação 12369 foi encontrada tanto na Califórnia quanto em Baltimore Maryland.

Tentou-se realizar na ferramenta T-COFFEE o mesmo procedimento adotado pelo CLUSTALW e MAFFT, mas infelizmente não houve sucesso, em função do limitador de entrada de dados. A ferramenta foi localizada em dois servidores web e ambos possuíam limite de entrada de dados de no máximo 10 mil caracteres. Considerando apenas a sequencia original e os mutantes encontrados no banco de dados do NCBI, a ferramenta deveria possuir um limite de entrada de dados superior a 40 mil caracteres (DEBIAN, 2014)

Em algumas consultas realizadas ao longo desta pesquisa nos bancos de dados do NCBI, EBI e DDBJ foi observado que constantemente são inseridas novas sequencias de mutantes e também de genes que ainda não estavam disponíveis nos bancos de dados. O volume de informações é sempre crescente não sendo possível afirmar que a quantidade de mutantes encontrados nesta pesquisa seja um valor fixo. Constantemente os bancos de dados com informações sobre genes estão sendo atualizados o que pode ser afirmado também por Edwards, Stajich e Hasen (2009).

O uso da Informática nos campos de estudo da vida é essencial para a interação entre as ciências, propiciando assim um melhor desenvolvimento humano e tecnológico. Nos últimos anos, a Biologia vem se apropriando com avidez das ferramentas proporcionadas pela Informática. Assim, a computação tornou-se peça chave nas pesquisas que vem sendo realizadas em áreas especializadas como a Biologia Molecular e o armazenamento de dados biológicos em bancos de dados públicos tem se tornado cada vez mais comum (SABBATINI, 1993).

Dados biológicos armazenados em banco de dados com acesso público, tais como, sequências de DNA e de proteínas tem facilitado o trabalho de

muitos cientistas, principalmente nos estudos de comparações de sequências de nucleotídeos para diversas finalidades, tais como: identificação de mutações ou polimorfismos genéticos, estudos evolutivos e filogenéticos. A disponibilidade dessas informações e a facilidade de acesso a internet, revolucionou a forma como as pesquisas, desta área, estão sendo realizadas na atualidade reduzindo significativamente o tempo destinado aos trabalhos de pesquisa relacionadas a sequenciamento genético (BAXEVANIS, 1998).

Os resultados permitiram afirmar que há a viabilidade de se estabelecer uma estratégia de análise genômica através das ferramentas propostas e que outros trabalhos poderão ser desenvolvidos com o foco em materiais sequenciados na região e comparados com dados de vários institutos de pesquisas localizados em diversos pontos do planeta, tal como preconizado no trabalho de Prosdocimi (2007).

# 6 CONCLUSÕES E PERSPECTIVAS FUTURAS

A internet possibilita que bancos de dados públicos de sequências de genoma ofereçam serviços por meio de uma interface uniforme para uma comunidade mundial de pesquisadores. Com um navegador web, um biólogo molecular pode, então, comparar uma sequência de DNA desconhecida com a coleção completa de sequências de DNA públicas. Na sequencia, pode disponibilizar seus resultados por meio da rede e contribuir com demais pesquisadores, tornando o volume de informações ainda maior.

A utilização destas metodologias viabilizou a realização de pesquisas com foco no estudo de exposição a agentes causadores de danos genéticos, de modo a disponibilizar técnicas avançadas para avaliação do grau de mutagenicidade apresentado por uma determinada ação antrópica ou natural. Proporcionou também a ampliação do conhecimento relacionado a pesquisa em banco de dados genômicos e das principais ferramentas de análise múltiplos de sequencias e estabeleceu condições para a implantação de uma estação de trabalho para realizar consultas a bancos de dados genômicos e comparações entre sequências de DNA.

Na Universidade do Oeste Paulista (UNOESTE), alguns grupos já produziram, por iniciativas isoladas, trabalhos de iniciação científica e monografias relacionadas com a Bioinformática (PUCCI NETO, 2001; OLIVEIRA, 2008). Entretanto, na atualidade não existem trabalhos de pesquisa em andamento nos cursos da Faculdade de Informática de Presidente Prudente – FIPP UNOESTE voltados para esta área, embora a disciplina de Bioinformática esteja presente na grade curricular de outros cursos de graduação da mesma universidade.

O conteúdo apresentado nesta pesquisa poderá contribuir para o caráter multidisciplinar destas atividades, aumentando a relevância e a produtividade nas análises Bioinformáticas. Neste contexto, a UNOESTE vem se preparando para uma inserção significativa na chamada era pós-genômica, de transformação da informação genômica em conhecimento científico e tecnológico.

Foi possível desenvolver um aprofundamento de estudos e experimentação que, usando os resultados de experimentos biológicos, possa alimentar análises de sequências de DNA, através das bases de dados acessíveis via Internet. A presença da estação de trabalho permitiu uma significativa elevação na eficiência de consulta às bases de dados disponíveis em diversos institutos de

pesquisas localizados em vários países.

Esta pesquisa consolidou uma boa base de conhecimento para a instalação de uma estação de trabalho de Bioinformática no laboratório de Citogenômica e Bioinformática da UNOESTE. Neste sentido, é importante ressaltar que, na região do Oeste Paulista está sendo implantado um hospital destinado ao tratamento de pacientes com canceres. Pode-se supor que, com a disponibilidade dos recursos e conhecimentos desenvolvidos nesta dissertação, as amostras biológicas dos pacientes poderão ser sequenciadas e analisadas de forma comparativa com sequencias gênicas mutantes oriundas de pacientes de vários países de modo a verificar eventuais semelhanças ou a identificação de características genômicas específicas dos mesmos.

A fim de transformar os dados desta pesquisa de forma mais dinâmica, trabalhos futuros poderão ter por objetivos a inserção das sequências mutantes em softwares de visualização da proteína a partir de suas sequencias de aminoácidos. A construção da proteína P53 em formato tridimensional possibilitará a análise dos seus sítios ativos e correlacionar as eventuais mutações com alterações no seu padrão de atividade.

## REFERÊNCIAS

ALTSCHUL, S. F.; MADDEN, T. L.; SCHÄFFER, A. A. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, Oxford, v. 25, n. 17, p. 3389-3402, jul. 1997. Disponível em: < http://nar.oxfordjournals.org/content/25/17/3389.full.pdf+html>. Acesso em: 5 jul 2014.

AMARAL M.A. et al. O programa BLAST: guia prático de utilização. Embrapa Recursos genéticos e Biotecnologia. Brasília, 2007. Disponível em: <https://www.embrapa.br/documents/1355163/2023605/doc224.pdf/ce77302f-b280-4c86-8857-bbe4b6bafb67>. Acesso em: ago 2014.

BAXEVANIS, A. D., OUELLETTE, B. F. F., **Bioinformatics**: a practical guide to the analysis of genes and proteins. [S.l.]: Hardcover, 2004.

BENSON, D. A. et al. GenBank**. Nucleic Acids Research**, Oxford, v. 27, n. 1, p. 12-17, nov. 2010. Disponível em: < http://nar.oxfordjournals.org/content/39/suppl_1/D32.full.pdf+html>. Acesso em: 10 ago 2014.

BOLLAG, G. et al. Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. **Nature,** New Jersey, v. 467 n.7315, p. 596-599, set. 2010,

CARVALHO, D. L. Metodologias para avaliação de impactos ambientais de aproveitamentos hidrelétricos. [S.l.}: [s.n.], 2010

CETESB. **Manual para elaboração de estudos para o licenciamento com avaliação de impacto ambiental**. São Paulo: CETESB, 2014. Disponível em: < http://www.cetesb.sp.gov.br/userfiles/file/dd/Manual-DD-217-14.pdf >. Acesso em: 20 jul 2014.

Japão Computational Biology Research Center (CBRC). Disponível em: < http://www.cbrc.jp/> Acesso em Nov 2014.

CORAPCIOGLU D. et al. Papillary micro carcinomas of the thyroid gland and immunohistochemical analysis of expression of p53 protein in papillary micro carcinomas. **Journal of Translational Medicine,** Ankara, v. 4, n. 1, p. 6, jul. 2006. Disponível em: < http://www.translational-medicine.com/content/pdf/1479-5876-4-28.pdf >. Acesso em: 25 abr 2014.

CHRISTOPHERSON, W. R. Geossistemas: uma introdução à geografia física. 7. ed. Porto Alegre: Bookman, 2012.

DEBIAN. **Pacote T-COFFE**: alinhamento de sequências múltiplas. Disponível em: https://packages.debian.org/pt/wheezy/t-coffee. Acesso em: 20 nov. 2014.

DESSLER, A**. The chemistry and Physics of stratospheric ozone.** Londres:. Academic Press, 2000.

EDWARDS, D.; STAJICH, J.; HASEN, D. **Bioinformatics:** tools and applications. New York: Springer, 2009.

EFEYAN, A.; SERRANO, M. p53: guardian of the genome and policeman of the oncogênese. **Cell Cycle,** Madrid, v. 2, n. 6, p. 1006-1010, mai. 2007. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/17457049>. Acesso em: 30 set 2014.

ESTADOS UNIDOS. National Center for Biotechnology Information. Disponível em: <http://www.ncbi.nlm.nih.gov/ >. Acesso em: 20 nov. 2014.

EUROPEAN MOLECULAR BIOLOGY LABORATORY.Disponível em: <http://www.ebi.ac.uk/ >. Acesso em: 20 nov. 2014a.

EUROPEAN MOLECULAR BIOLOGY LABORATORY. **CLUSTALW**. Disponível em: <http://www.ebi.ac.uk/Tools/msa/CLUSTALW2/help/>. Acesso em: 20 nov. 2014b.

EUROPEAN MOLECULAR BIOLOGY LABORATORY. **MAFFT**. Disponível em: <http://www.ebi.ac.uk/Tools/msa/mafft/help/>. Acesso em: 15 nov. 2014c.

EUROPEAN MOLECULAR BIOLOGY LABORATORY. **T-COFFEE**. Disponível em: <http://www.ebi.ac.uk/Tools/msa/tcoffee/help/>. Acesso em: 15 nov. 2014d.

FREBOURG, T. et al. Germ-line mutations of the p53 tumor suppressor gene in patients with high risk for cancer inactivate the p53 protein. **Proceedings of the National Academy of Science of the United State of America**, v. 89, p. 6413-6417, jul. 1992. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC49511/pdf/pnas01088-0196.pdf>. Acesso em: 01 nov 2014.

GERAQUE, E. A revolução da bioinformática. **Revista Pesquisa Fapesp.** São Paulo, v. 86, n. 55, p. 6413-6417, abr., 2003. Disponível em: <http://revistapesquisa.fapesp.br/2003/04/01/a-revolucao-da-bioinformatica/>. Acesso em: 01 out. 2014.

GIBAS, C.; JAMBECK, P., **Desenvolvendo bioinformática**: ferramentas de software para aplicações em biologia. Rio de Janeiro: Campus, 2001.

HAINAUT, P.; HOLLSTEIN, M. p53 and human cancer: the first ten thousand mutations. **Adv. Câncer Res.**, Lyon, v. 77, p. 81–137, 2000. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/10549356>. Acesso em: 5 ago 2014.

INSTITUTO NACIONAL DE CÂNCER (BRASIL). **Coordenação de prevenção e vigilância estimativa 2014:** incidência de câncer no Brasil. Rio de Janeiro: INCA, 2014. Disponível em: <http://www.inca.gov.br/estimativa/2014/estimativa-24042014.pdf> Acesso em: 10 nov. 2014.

KARP, Gerald. **Biologia celular e molecular**: conceitos e experimentos. Barueri: Manole, 2005.

KLEIHUES, P. et al. Tumors associated with p53 germline mutations. **American**

**Journal of Pathology,** Lyon, v. 150, p. 1-13, 1997:. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1858532/pdf/amjpathol00025-007.pdf>. Acesso em: 10 out 2014.

LIMA, D. S. **Estratégia paralela exata para o alinhamento múltiplo de sequências biológicas utilizando unidades de processamento gráfico (GPU)**. 2012. 73 f. Dissertação (Mestrado em Informática) - Universidade de Brasília, Brasília.

MAGNONI, M. S.; FRANCISCO, O. Expressão da proteína p53 na reparação do DNA correlacionando oncogêneses em humanos. In: CIC - CONGRESSO DE INICIAÇÃO CIENTÍFICA, 8., 2009, Ourinhos. **Anais...** Ourinhos: Faculdades Integradas de Ourinhos, 2009. Disponível em: <http://fio.edu.br/cic/anais/2009_viii_cic/autores.html>. Acesso em: 10 jan 2015.

MAPA dos Estados Unidos desenhado em 4 cores. In: Wikpédia: a enciclopédia livre. 2014. Disponível: < https://pt.wikipedia.org/wiki/Teorema_das_quatro_cores#/media/File:Map_of_USA_with_state_names_pt.svg> . Acesso em: out 2014.

MARTINEZ, M. A. R. et al. Genética molecular aplicada ao câncer cutâneo não melanoma. **Educação Médica Continuada,** São Paulo, v. 81, n. 05, p. 405-19, 2006. Disponível em: <http://www.scielo.br/pdf/abd/v81n5/v81n05a03.pdf>. Acesso em: 15 jun 2014.

MELO, Marciano Almeida. o desenvolvimento industrial e o impacto no meio ambiente. 2012. **E-gov.** Disponível em: <http://www.egov.ufsc.br/portal/conteudo/o-desenvolvimento-industrial-e-o-impacto-no-meio-ambiente> Acesso em: nov. 2014.

MORITA, N.; IKEDA, Y.; TAKAMI, H.. Clinical significance of p53 protein expression in papillary thyroid carcinoma. **World J. Surg.,** Tokyo, v. 32, p. 2617-2622, 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/18836853>. Acesso em: 30 mai 2014.

OLIVEIRA V. S. J. **Análise comparativa de similaridade de sequências de ácidos nucleicos**. 2008. 65 f. Monografia de Especialização (Ciências Biológicas) - Universidade do Oeste Paulista, Presidente Prudente.

PARKIN, D. M. International variation. **Oncogênese,** Lyon, v. 23, p. 6329-6340, 2004. Disponível em: <http://www.nature.com/onc/journal/v23/n38/full/1207726a.html>. Acesso em: 5 mai 2014.

PROSDOCIMI F. et al. Bioinformática: manual do usuário. **Biotec. Ci. Des**, n. 29, p.18-31, 2002. Disponível em: <http://www.academia.edu/546335/Bioinform%C3%A1tica_manual_do_usu%C3%A1rio>. Acesso em: 25 out 2014.

PROSDOCIMI, F. **Curso on line**: introdução a bioinformática: biotecnologia ciência & desenvolvimento. Brasília: Portal Biotecnologia, 2007. Disponível em:

<http://biotec.icb.ufmg.br/chicopros/Prosdocimi07_CursoBioinfo.pdf>. Acesso em: 10 ago 2014.

PUCCI NETO J. Comparação de sequências de DNA. 2001. 43 f. Monografia de Graduação (Ciência da Computação) - Universidade do Oeste Paulista, Presidente Prudente.

RIGDEN, D. J.; MELLO, L. V. Anotação funcional computacional de proteínas: novos métodos computacionais poderão preencher lacunas do sistema de anotação atual. **Biotecnologia Ciência & Desenvolvimento**, São Paulo, v. 5, n. 29, nov./dez. 2002. Disponível em: <http://www.biotecnologia.com.br/revista/bio25/anota.pdf>. Acesso em: 20 nov 2014.

ROCHA, J. C.; ROSA, A. H.; ARNALDO, A. **Introdução à química ambiental**. 2. ed. Porto Alegre: Bookman, 2009.

SABBATINI, R. M. E. Informática e biologia: a nova aliança. In: SIMPÓSIO DE APLICAÇÕES DA INFORMÁTICA EM BIOLOGIA, 1., 1993, Campinas. **Anais...** Campinas: UNICAMP, 1993. Disponível em: <http://www.renato.sabbatini.com/papers/I%20SIMPOSIO%20DE%20APLICACOES%20DA%20INFORMATICA%20EM%20BIOLOGIA.pdf>. Acesso em: 5 mai 2014.

SANCHEZ, L. E. **A efetividade da avaliação de impacto ambiental no estado de São Paulo:** uma análise a partir de estudos de caso. São Paulo: Secretaria do Meio Ambiente, Coordenadoria de Planejamento Ambiental, 1995. p. 13-19.

SANSOM, C. E.; SMITH, C. A Computer applications in biomolecular sciences: part 2: bioinformatics and genome projects**. Biochemical Education,** Manchester, v. 28, p. 127-131, 2000. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.1539-3429.2000.tb00043.x/epdf>. Acesso em: 30 set 2014.

SANTOS, Eduardo Campos dos. **Uma introdução à bioinformática através da análise de algumas ferramentas de software livre ou de código aberto utilizadas para o estudo de alinhamento de sequências**. 2004. 87 f. Monografia (Especialização em Administração em Redes Linux) - Universidade Federal de Lavras.

SANTOS, Elaine Teresinha Azevedo dos. **Educação ambiental na escola:** conscientização da necessidade de proteção da camada de ozônio. 2007. 52 f. Monografia (Especialização em Educação Ambiental) - Universidade Federal de Santa Maria.

SETUBAL, J. C.; MEIDANIS, J. **Introduction to computational molecular biology.** Campinas: UNICAMP, 1997.

SIMÃO T. A. et al. TP53 mutations in breast cancer tumors of patients from Rio de Janeiro, Brazil: association with risk factors and tumor characteristics. **Int. J. Cancer**. São Paulo, v. 101, n. 1, p. 69-73, 2002. Disponível em: <http://www.scielo.br/pdf/rbepid/v7n2/13.pdf>. Acesso em: 30 set 2014.

SOUZA V.J. **Projeto genoma humano**. São Paulo: Ed. Loyola, 2004.

STRAUSS, B. S. Silent and multiple mutations in p53 and the question of the hypermutability of tumors. **Carcinogenesis,** Chicago, v. 18, p. 1445-1452, 1997. Disponível em: <http://carcin.oxfordjournals.org/content/18/8/1445.full.pdf>. Acesso em: 10 ago 2014.

THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. CLUSTAL W: improving the sensiti-vity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Research**, Heldelberg, v. 22, n. 22, p. 4673-4680, nov. 1994. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC308517/pdf/nar00046-0131.pdf>. Acesso em: 20 out 2014

TICONA, W. G. C. **Aplicação de algoritmos genéticos multi-objetivo para alinhamento de sequências biológicas**. 2003. 129 f. Dissertação (Mestrado em Ciências) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

VARLEY, J.M., et al. Characterization of germline TP53 splicing mutations and their genetic and functional analysis. **Oncogene,** Manchester, v. 20, p. 2647-2654, 2001.Disponível em: <http://www.nature.com/onc/journal/v20/n21/pdf/1204369a.pdf>. Acesso em: 05 nov 2014.

WATSON, J.D., et al. **Biologia molecular do gene**. Porto Alegre: Artmed, 2006.

ZHANG, J.; MASSEN, T. L. **PowerBLAST**: a new network BLAST application for interactive or automated sequence analysis and annotation. National Center for Biotechnology Information (NCBI), Bethesda, 1997. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC310664/>. Acesso em: 10 dez 2014.