UNIVERSIDADE FEDERAL DE ALFENAS
INSTITUTO DE CIÊNCIAS EXATAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

INTEGRAÇÃO DE FERRAMENTAS WEB PARA ANÁLISE DE DADOS GENÔMICOS E PROTEÔMICOS.

Bruno Augusto Terra
Lucas Francisco de Moura

Orientador: Prof. Dr. Nelson José Freitas da Silveira

Alfenas, MG
2017

UNIVERSIDADE FEDERAL DE ALFENAS
INSTITUTO DE CIÊNCIAS EXATAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

INTEGRAÇÃO DE FERRAMENTAS WEB PARA ANÁLISE DE DADOS GENÔMICOS E PROTEÔMICOS.

Bruno Augusto Terra

Lucas Francisco de Moura

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal de Alfenas como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Nelson José Freitas da Silveira.

Alfenas, MG
2017

Bruno Augusto Terra
Lucas Francisco de Moura

INTEGRAÇÃO DE FERRAMENTAS WEB PARA ANÁLISE DE DADOS GENÔMICOS E PROTEÔMICOS.

A Banca examinadora abaixo-assinada aprova a monografia apresentada como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação pela Universidade Federal de Alfenas.

Profª.Dª. Marcia Paranho Veloso

Universidade Federal de Alfenas

Profª. Poliany Graziella de Freitas

Universidade Federal de Alfenas

Mᵉ. Thiago Castilho Elias

Universidade Federal de Alfenas

Alfenas, MG
2017

AGRADECIMENTOS

Agradecemos primeiramente a Deus pelo dom da vida e por nos proporcionar chegar até aqui.

Agradecemos a nossos familiares e amigos pelo apoio, dedicação e paciência que nos foi dado durante toda essa jornada.

Agradecemos aos professores que sempre estiveram dispostos a ajudar e contribuir para o melhor aprendizado, em especial ao orientador Prof. Dr. Nelson José Freitas da Silveira, que nos auxiliou no desenvolvimento deste trabalho.

Agradecemos também a Universidade Federal de Alfenas- UNIFAL, por ter nos dado a chance e todas as ferramentas que permitiram chegar hoje ao final deste curso de maneira satisfatória.

A todos que direta ou indiretamente fizeram parte da nossa formação, o nosso muito obrigado.

RESUMO

Diversas técnicas e programas computacionais auxiliam o estudo da estrutura dos genomas. Por meio da bioinformática é possível realizar a identificação de proteínas, análises de imagens, de localização, função e interação. O uso da tecnologia além de proporcionar economia para as empresas, aumenta a confiabilidade e assegura registros dos resultados, reduz o tempo do processamento de dados e diminui a taxa de erros. Este trabalho teve como objetivo desenvolver uma plataforma web, a EMGPA(nome decorrente da letra inicial de cada programa), na linguagem Php para integração dos softwares Emboss, Modeller, Pipegen, Autodock e Gramm, organizando grandes extensões de dados que são gerados em pesquisas de maneira simplificada e acessível, possibilitando a análise por um maior número possível de cientistas. A utilização da plataforma web de integração permite um melhor gerenciamento das informações de pesquisas, otimizando o tempo e obtendo maior produtividade.

Palavras-Chave: Plataforma web, integração de softwares, genômica, proteômica, bioinformática.

ABSTRACT

Several techniques and computational programs help to study the structure of genomes. Through bioinformatics it is possible to perform protein identification, image analysis, location, function and interaction. Besides, the use of technology to provide economy for businesses, increases reliability and ensures records of results, reduces data processing time and reduces error rate. This project aims to develop a web platform, an EMPGA(succinct name of the initial letter of each program), in the PhP language for integration of Emboss, Modeller, Pipegen, Autodock and Gramm software organizing large extensions of data that are generated in researches in a simplified and accessible way and enabling the analysis by as many scientists as possible. The use of the integration web platform allows a better management of the search information, optimizing the time and obtaining greater productivity.

Keywords: Web platform, software integration, genomics, proteomics, bioinformatics.

Lista de Figuras

Figura 1 - Página Inicial.....30
Figura 2 - Página de acesso aos programas.....31
Figura 3 - Página Needle.....31
Figura 4 - Ajuda do Needle.....32
Figura 5 - Download Resultado.....33

SUMÁRIO

1. INTRODUÇÃO 14
2. OBJETIVO GERAL 18
2.1 OBJETIVOS ESPECÍFICOS 18
3. JUSTIFICATIVA 19
4. REFERENCIAL TEÓRICO 20
4.1 IMPORTÂNCIA DA UTILIZAÇÃO DE SOFTWARES EM LABORATÓRIOS 20
4.2 EMBOSS (European Molecular Biology Open Software Suite) 20
4.3 MODELLER 21
4.4 AUTODOCK 22
4.5 GRAMM (Global Range Molecular Matching) 23
4.6 PIPEGEN 24
4.7 PHP (Hypertext Preprocessor) 25
5. MATERIAIS E MÉTODOS 27
6.RESULTADOS E DISCUSSÕES 30
7. CONCLUSÃO 34
8. REFERÊNCIAS BIBLIOGRÁFICAS 35

1. INTRODUÇÃO

Cada organismo apresenta individualidades, essas características podem ser determinadas quando analisamos suas sequências genômicas. Os estudos do sequenciamento genômico iniciaram-se em 1995, quando a bactéria *Haemophilus influenzae* teve seu genoma sequenciado pela estratégia de sequenciamento genômico completo por fragmentos aleatórios de DNA, realizado por um grupo de pesquisadores TIRG (*The Institute for Genomic Research – O Instituto de Pesquisa genômica*) nos Estados Unidos (FLEISCHMANN et al., 1995).

Estes estudos de sequenciamento desenvolveram-se rapidamente e atualmente encontram-se disponíveis as sequências de genes de organismos inteiros, como plantas, animais e até mesmo o genoma humano. A Genômica consiste na ciência que estuda o conjunto de todo o DNA de um organismo (genoma). Um dos enfoques da pesquisa genômica é a construção de mapas detalhados de cromossomos (SNUSTAD; SIMONS, 2001).

Desde que Gregor Mendel realizou o experimento utilizando ervilhas e definiu os primeiros conceitos na genética, porém passou mais de um século para que os primeiros genomas viessem a ser sequenciados (SANGER et al., 1977).

Com a corrida pelo genoma humano no final da década de noventa e começo do novo século (DAVIES, 2001), novas metodologias passaram a ser traçadas para a produção de sequências genômicas de forma ainda mais massiva do que utilizando os sequenciadores de eletroforese capilar baseados em dideoxinucleotídeos fluorescentes (BOYSEN et al., 1997).

Empresas biotecnológicas e mentes inovadoras rapidamente perceberam que o surgimento de uma tecnologia de sequenciamento mais veloz certamente proporcionaria um salto em nossa compreensão sobre os genomas de diferentes organismos. (STEINDORFF, 2016).

Uma vez determinado o genoma, dá-se início à proteômica, que é o estudo do conjunto de proteínas produzidas por parte de um genoma. O estudo proteômico visa à identificação do conjunto de proteínas sintetizadas por uma determinada célula, tecido ou organismo durante um evento fisiológico específico como diferenciação, resposta a drogas ou transformação em células cancerosas. Tal

estudo permite também informações sobre a concentração protéica, modificações pós-traducionais e eventos de *splicing* (PANDEY; MANN, 2000).

O proteoma determina as proteínas expressas em um genoma ou tecido, enquanto o genoma indica a soma de todos os genes de um indivíduo, o proteoma não é uma característica variável de um organismo. O proteoma altera com o estado de desenvolvimento, do tecido ou mesmo sob as condições nas quais o indivíduo se encontra (GALDOS, 2009). Portanto, há muito mais proteínas no proteoma do que genes no genoma, especialmente para eucariotos (ROCHA et al., 2005).

O estudo das proteínas não abrange apenas a soma dos produtos traduzidos a partir das sequências genômicas, mas inclui também proteínas resultantes de processos pós-transcricionais e pós-traducionais, bem como complexos formados por essas biomoléculas (AHRENS et al., 2010). Além de toda a complexidade envolvida no proteoma, ele possui um perfil dinâmico e que se altera de acordo com o status fisiológico e as fases da diferenciação celular. Algumas estimativas sugerem que mais de um milhão de diferentes tipos de proteínas estão presentes nas células, nos tecidos e nos fluidos corporais em condições e/ou momentos distintos (JENSEN et al., 2004).

O termo proteômica refere-se ao estudo do conjunto dessas moléculas, que são responsáveis direta ou indiretamente pelo controle de todos ou quase todos os processos biológicos. A proteômica estuda de forma descritiva e quantitativa desde o conjunto de proteínas de uma organela subcelular até aquelas de um ecossistema, suas variações na população, mudanças em resposta a um ambiente ou decorrentes do desenvolvimento normal ou alterado, e modificações e interações com outras proteínas (VALLEDOR; JORRIN, 2011).

A proteômica pode ser dividida em três áreas principais: Microcaracterização protéica para a identificação em larga escala de proteínas e suas modificações pós-traducionais; determina a diferença para a comparação dos níveis de proteína com aplicação potencial em uma ampla faixa de doenças; e estudos das interações proteína-proteína utilizando técnicas tais como a Espectrometria de Massas (PANDEY; MANN, 2000). A separação de proteínas por técnicas bidimensionais como ponto isoelétrico seguido de eletroforese em gel de poliacrilamida (mapeamento proteico) é um método que tem apresentado características marcantes, pois apresentam alto poder de resolução e boa reprodutibilidade quando são usados para separar soluções complexas de proteína, assim como extratos de plantas ou soro humano (KLOSE, 1975). Atualmente a eletroforese bidimensional (2D) é a principal

técnica de separação de proteínas utilizada antes da aplicação da amostra no espectrômetro de massas.

Sua vantagem em relação a outras tecnologias é a capacidade de separar com alta resolução um grande número de proteínas de uma amostra complexa e a possibilidade de se fazer análise de expressão gênica por meio de comparação dos padrões proteicos (QUALTIERI et al., 2007; SIZOVA et al., 2007). Para separação de proteínas por eletroforese uni (1-DE) e bidimensional (2-DE), deve se isolar as moléculas de materiais biológicos, tais como tecidos e fluidos corporais. Para a obtenção de bons resultados é fundamental a extração adequada de proteínas. Em função da variedade de tipos e origens de amostras biológicas, o procedimento de extração necessita de otimização individual. Na maioria dos casos, as proteínas precisam ser solubilizadas, desagregadas, desnaturadas e submetidas a tratamento com agentes redutores de pontes dissulfeto (MARQUI et al., 2006).

A preparação apropriada das amostras é essencial para obter bons resultados nos estudos da proteômica. O melhor método de extração, precipitação e solubilização das proteínas varia de uma amostra para outra e deverá ser estabelecido para cada caso em particular (RABILLOUD, 1996; CORTHALS et al., 2000; STULTS; ARNOTT, 2005; GORG et al., 2010; USAMI et al., 2007; GALDOS, 2009). A análise das imagens do gel utilizando softwares especializados permite comparações de múltiplos géis e, através da interligação com bases de dados detalhadas do proteoma na internet. Assim, por um processo de subtração, as diferenças (por exemplo, presença, ausência, ou intensidade das proteínas) entre amostras sadias e doentes podem ser reveladas (RIVEROS et al., 2010).

Diversas técnicas e programas computacionais auxiliam o estudo da estrutura dos genomas. Por meio da bioinformática é possível realizar a identificação de proteínas, análises de imagens, de localização, função e interação. Existem softwares capazes de capturar imagens provenientes de géis resultantes de experimentos de eletroforese e fazem nele a varredura das proteínas apresentadas nas imagens sob a forma de spots (SILVA, 2011).

Porém como tratamos de experimentos complexos com uma grande extensão de dados, muitas vezes a integração de softwares auxilia neste processo, tornando-se fundamental em projetos de análises proteômicas, permitindo que os dados obtidos possam ser rastreados durante todo o experimento.

A integração dos softwares Emboss, Pipegen, Autodock, Gramm, Modeller, os quais foram utilizados neste projeto permitiram uma análise e modelagem de proteínas e sequência de DNA.

2. OBJETIVO GERAL

Desenvolver uma plataforma web(EMGPA) a fim de auxiliar os pesquisadores que trabalham com sequenciamento genético e proteômico. Por meio do recebimento de dados oriundos de softwares de bioinformática estes serão analisados, correlacionados e serão produzidos relatórios.

2.1 OBJETIVOS ESPECÍFICOS

O projeto tem como objetivos específicos:

- Construir uma plataforma web utilizando a linguagem de programação php.
- Permitir que a plataforma seja capaz de incorporar dados de pesquisa.
- Auxiliar por meio do software pesquisadores a analisar os experimentos.

3. JUSTIFICATIVA

A utilização de softwares para análises de sequenciamento é fundamental para pesquisadores, para facilitar à análise de dados oriundos de pesquisas, capacitando a integração de informações. Portanto, este trabalho busca integrar softwares para receber e relacionar dados que irão auxiliar a compreensão dos resultados recebidos em experimentos genéticos.

4. REFERENCIAL TEÓRICO

Várias ferramentas desenvolvidas pela bioinformática permitem o acesso e análise de dados no GenBank. Com o início do projeto Genoma Humano em 1990 e subsequente disponibilização de sequenciadores automáticos de DNA capazes de gerar dados genômicos em grande escala, os bancos de dados e ferramentas de análises tiveram que se adaptar a este volume crescente de informações, Sequências de nucleotídeos são adicionadas aos bancos de dados (como o GenBank) na ordem de milhares de pares de bases (pb) por segundo todos os dias. Nos serviços de bioinformática de projetos genoma, essas inúmeras sequencias individuais, cada uma portando geralmente entre 400 a 1000 pb, devem ser montadas em sequencias cada vez maiores, os *contigs*, através de ferramentas que avaliam a qualidade das sequencias individuais e a superposição destas, para que finalmente sejam disponibilizados segmentos cromossômicos inteiros de alta qualidade. Para a cobertura total de um genoma com boa qualidade estima- se que este deva ser sequenciado ao equivalente a dez vezes seu tamanho em pares de bases (VENTER et al., 2001).

4.1 IMPORTÂNCIA DA UTILIZAÇÃO DE SOFTWARES EM LABORATÓRIOS

Atualmente os laboratórios cada vez mais utilizam a tecnologia para melhorar e facilitar suas análises. O uso da tecnologia, além de proporcionar economia para as empresas, aumenta a confiabilidade dos resultados e assegura registros dos mesmos, reduz o tempo do processamento de dados e reduz a taxa de erros.

A UNICAMP foi a pioneira no Brasil a desenvolver e aplicar pesquisas genômicas, o laboratório de bioinformática foi responsável pela montagem do genoma do primeiro organismo sequenciado no país em 2000, a bactéria *Xyllela fastidiosa,* causadora da doença amarelinho na laranja (SIMPSON et al., 2000).

4.2 EMBOSS (European Molecular Biology Open Software Suite)

EMBOSS (European Molecular Biology Open Software Suite) é um pacote gratuito desenvolvido para atender as necessidades da comunidade da biologia

molecular. O software lidar com dados em uma variedade de formatos, e permite a recuperação de dados e sequências a partir da web. O EMBOSS integra uma gama de pacotes e ferramentas disponíveis para análise de sequências. Ele quebra a tendência histórica para pacotes de software comerciais, abrangente de serviços integrados e de aplicações de análise de sequências lançado como código aberto na plataforma UNIX, serve tanto para análise de sequências como para o desenvolvimento de programas de análise (RICE et al., 2000).

O EMBOSS tende a ser difícil de usar e nem sempre interage perfeitamente com outros programas. Pacotes completos de aplicativos integrados de fornecedores comerciais são mais fáceis de usar, mas o valor pode ser muito alto para aquisição em laboratórios e não podem ser customizados para uso local porque o código-fonte não está disponível.

Os criadores do EMBOSS objetivaram–se em proporcionar a biologia molecular abrangente de aplicações de análise de seqüência integrada lançado sob o modelo de código aberto em plataformas UNIX. O EMBOSS é um software aberto projetado com dois objetivos: Para fornecer à comunidade de biologia molecular um pacote de software livremente disponível para analisar seqüências de DNA e proteínas e fornecer um conjunto de bibliotecas de software para que os cientistas usem para desenvolver suas próprias aplicações (RICE et al., 2000). O EMBOSS geralmente é utilizado para alinhamento de sequenciamento genético, pesquisa rápida em banco de dados com padrões de sequência, análise de padrões de sequência de nucleotídeos, rápida identificação de padrões de sequência em conjuntos de sequências de grande escala.

Como o EMBOSS possui uma série de pacotes, à plataforma EMGPA serão implementados os principais pacotes: Needle, Stretcher, Water, Matcher, Transeq, Pepinfo e Seqret.

4.3 MODELLER

O Modeller é um programa da bioinformática utilizado para modelagem de proteínas por homologia, que descreve patamares metabólicos e seus componentes: reações, enzimas e substratos. Extremamente popular e eficiente de modelagem comparativa, os comando para a utilização devem ser inseridos através de linhas de

comando num terminal, exigindo prévio conhecimento de informática para utiliza-lo (SALI et al., 1995).

Ele é capaz de modelar estruturas de proteínas tridimensionais e a monta de acordo com a técnica de satisfação de restrições espaciais, o programa é usado para homologia ou para comparar modelos de estruturas de proteínas, o usuário fornece um alinhamento de uma sequência a ser modelado com estruturas relacionadas conhecidas e o Modeller automaticamente calcula um modelo com todos os átomos que não sejam hidrogênio.

A Modelagem por homologia, nada mais é que o mecanismo evolucionário de duplicação de genes, associado às mutações, leva a divergências moleculares e, consequentemente, à formação de famílias de proteínas estruturalmente relacionadas. Proteínas derivadas de um ancestral comum são ditas homólogas. Para identificarmos à modelagem homóloga é necessário 4 passos, são eles: identificação e seleção de proteínas-molde, alinhamento de sequencias de resíduos, construção de coordenadas do modelo e validação.

Com o desenvolvimento de técnicas e de estudos do genoma a utilização de softwares tornou-se indispensável. O Modeller contém em seu banco de dados relacional modelos para segmentos de mais de 300 proteínas identificadas no genoma (SILVEIRA et al., 2005).

O software também é muito utilizado para o desenvolvimento de fármacos, permitindo identificar a localização do local de ligação das moléculas e a geometria da ligação nos sítios ativos (AZEVEDO et al., 1996). Portanto, esta base de dados que possui como objetivo principal a visualização da estrutura de proteínas, ainda é capaz de auxiliar na área da saúde para o desenvolvimento de fármacos.

4.4 AUTODOCK

O AutoDock é um software gratuito, amplamente distribuído para fins acadêmicos e comerciais, já foi distribuído para mais de 2.9000 usuários em todo o mundo, de acordo com pesquisas realizadas em janeiro de 2011, uma pesquisa do ISI Citation Index mostrou que mais de 2700 publicações citam os métodos AutoDock como ferramenta fundamental.

AutoDock é um conjunto de ferramentas automatizadas para docking molecular, é utilizado para prever como moléculas pequenas, tais como substratos ou candidatos a possíveis fármacos, se ligam a um receptor de estrutura 3d conhecida (MORRIS, 2016)

O software apresenta-se em duas gerações disponíveis, o AutoDock 4 e o AutoDock Vina. O AutoDock 4 executa o *acoplamento* do docking para um conjunto de grades que descrevem a proteína do alvo; *Autogrid* pré-calcula essas grades. Portanto além de utilizá-los para encaixe, as redes de afinidade atômica podem ser visualizadas. Isto pode ajudar, por exemplo, a orientar os químicos orgânicos sintéticos a conceberem melhor aglutinantes.

AutoDock Vina lançado posteriormente, não requer a escolha de tipos de átomos e mapas de grade de pré-cálculo para eles. Em vez disso, ele calcula as grades internamente, para os tipos de átomo que são necessários, fazendo isso praticamente instantaneamente (HOSOUME, 2016).

O AutoDock Vina foi concebido em 2009 e foi completamente reestruturado. Versões anteriores do AutoDock, como o AutoDock 4, utilizam outros algoritmos de busca, entre eles o algoritmo genético lamarquiano (lamarckian genetic algorithm) para a procura da melhor posição, orientação e rotação do ligante. O AutoDock Vina, para encontrar o melhor modo de ligação, emprega uma técnica estocástica denominada Local Search global optimizer. Nessa técnica são repetidos diversos passos que envolvem mutação e otimização local, de modo que o resultado encontrado é aceito conforme o critério de Metropolis (HOSOUME, 2016).

Atualmente é empregado para realização de análises de estruturas de fármacos, cristalografias de raio X, triagens virtuais, estrutura de biblioteca combinatória, acoplamento de proteína, estudos de mecanismos químicos, entre outros. É considerado um software muito rápido e eficiente, já que fornece alta qualidade nos resultados das previsões experimentais (HOSOUME, 2016).

4.5 GRAMM (Global Range Molecular Matching)

Devido ao aumento das necessidades de biologia experimental acarretou no desenvolvimento de tecnologias computacionais confiáveis para suprir as carências na modelagem de interações protéicas. O progresso recente em algoritmos de

ancoragem e hardware de computador torna possível implementar procedimentos como servidores web automatizados, o que melhora consideravelmente a utilidade das abordagens de ancoragem na comunidade biológica. Uma característica importante do Gramm é a capacidade de suavizar a representação da superfície da proteína para ter em conta a possível alteração conformacional na ligação dentro da abordagem de ancoragem de corpo rígido. A simplicidade da interface e instalação, bem como a sua disponibilidade na plataforma Windows são outros pontos fortes que contribuem para a popularidade deste programa por pesquisadores. (TOVCHIGRECHKO; VAKSER, 2006).

Gramm é um programa para o acoplamento de proteínas, para prever a estrutura de um conjunto de proteínas que necessita somente das coordenadas atômicas das duas moléculas, sem a necessidade de se conhecer informações sobre as suas ligações. Ele realiza uma pesquisa 6-dimensional exaustiva através das traduções relativas e rotações das moléculas. Os pares moleculares podem ser: duas proteínas, uma proteína e um composto menor, duas hélices transmembranares, etc. O Gramm pode ser usado para moléculas de alta resolução, para estruturas imprecisas (onde apenas são conhecidas as características estruturais grosseiras), em casos de grandes Mudanças conformacionais, etc. (VICKLUND, 2017).

A metodologia Global Range Molecular Matching (GRAMM) é uma abordagem empírica para suavizar a função de energia intermolecular. A qualidade é determinada de acordo com a precisão das estruturas. Assim, o encaixe de estruturas de alta resolução com pequenas mudanças conformacionais produz uma previsão exata, enquanto o encaixe de estruturas de resolução ultra-baixa dará apenas as características brutas do complexo (VICKLUND, 2017).

4.6 PIPEGEN

O PipeGen foi desenvolvido no intuito de facilitar a transferência de dados entre bancos de dados, onde o usuário possui várias análises que partem de programas diferentes e utilizando o PIPEGEN ele combina esses dados para poder obter um resultado final. Na plataforma será utilizado para integração dos resultados dos softwares.

Ao utilizar uma combinação de análise de código estático e dinâmico, PipeGen automaticamente incorpora um pipe de dados em um SGBD (sistema de gerenciamento de banco de dados). Aproveitando do fato de que muitos SGBDs podem exportar e importar dados a partir de formatos comumente usados (por exemplo, CSV), e que a maioria dos sistemas vêm com testes de unidade que exercem essa funcionalidade. PipeGen leva uma série de insumos, incluindo um conjunto desses testes unitários e a fonte código do SGBD. (HAYNES et al., 2016)

PipeGen é capaz de analisar o código fonte do banco de dados e executar cada um dos testes de unidade de exportação para criar um pipe de dados que redireciona os dados exportados do disco para um soquete de rede que é fornecido pelo PipeGen em tempo de execução. (HAYNES et al., 2016)

O PipeGen estende as implementações de formato de dados orientadas por texto com código que intercepta os dados originais antes de sua codificação de texto, identifica e elimina delimitadores e redundantes metadados. Ao capturar os dados originais, o PipeGen permite transmissão de dados usando um formato binário eficiente (Apache Arrow).

4.7 PHP (Hypertext Preprocessor)

O Php foi criado a partir da necessidade de se desenvolver sites web dinâmicos. Criada em 1994 por Rasmus Lerdorf, a linguagem Php é derivada da linguagem Perl, e, desde 1998, em sua terceira versão, o Php pode concorrer diretamente com as linguagens similiares, como Asp da Microsoft e o Jsp da Sun. O poder do Php está no fato dele ser uma linguagem interpretada, por trabalhar diretamente com o código fonte ao ser executado, diferente das linguagens compiladas, que necessitam de um arquivo binário como, por exemplo, os .exe (SAVOIA, 2013).

Um programa Php pode ser escrito em qualquer editor de texto. Já existem também diversos editores específicos para o Php, que exibem cada elemento (variáveis, textos, palavras reservadas etc.) com cores diferentes, para melhorar a visualização. Um trecho de código Php deve estar entre as tags < ?php ?> , para que o servidor Web possa reconhecer que trata-se de um código de programação e possa chamar o interpretador Php para executá-lo (NIEDERAUER, 2013).

Normalmente uma página Php não contém apenas códigos de programação Php, mas também tags de marcação Html. Enquanto o Php representa a parte dinâmica da página, a Html representa a parte estática (NIEDERAUER, 2013).

Essa combinação entre Html e php é muito útil, pois nós utilizamos o Php para gerar os dados de forma dinâmica, enquanto que o Html é utilizado para formatar e exibir esses dados nas páginas mostradas no navegador (NIEDERAUER, 2013). Ao longo de mais de uma década, o Php vem adicionando mais e mais recursos e se consolida ano após ano como uma das linguagens de programação orientada a objetos que mais crescem no mundo. Estima-se que o Php seja utilizado em mais de 80% dos servidores web existentes, tornando-a disparadamente a linguagem mais utilizada para desenvolvimento web (DALL'OGLIO, 2015).

5. MATERIAIS E MÉTODOS

Por meio de pesquisa em livros, artigos e relatos de pesquisadores da UNIFAL-MG (Universidade Federal de Alfenas) que relacionam- se com a área de biotecnologia, biologia molecular, genética, foi realizada a integração de softwares de acordo com as principais necessidades avaliadas. Atualmente temos diversos softwares destinados a analisar resultados em laboratórios, porém sistemas capazes de integrar resultados de diferentes softwares ainda são escassos. Foram realizadas entrevistas com pesquisadores envolvidos e verificou-se a necessidade de integrar os seguintes softwares: Emboss, Modeller, Pipegen, Autodock, Gramm.

Devido a essa imensa quantidade de dados gerados em inúmeros laboratórios de todo o mundo, faz-se necessário organizá-los de maneira acessível, de modo a evitar redundância na pesquisa cientifica e possibilitar a análise por um maior número possível de cientistas. A construção de banco de dados para armazenamento de informação de sequencias de DNA e genomas inteiros, proteínas e suas estruturas tridimensionais, bem como vários outros produtos da era genômica, tem sido um grande desafio, mas simultaneamente extremamente importantes.

Atualmente a bioinformática é imprescindível para a manipulação de dados biológicos. Ela pode ser definida como uma modalidade que abrange todos os aspectos de aquisição, processamento, armazenamento, distribuição, análise e interpretação da informação biológica. Através da combinação de procedimentos e técnicas da matemática, estatística e ciências da computação, são elaboradas várias ferramentas que nos auxiliam a compreender os significados biológicos representados nos dados genômicos. Além disso, com a criação de bancos de dados as informações já processadas aceleram a investigação em outras áreas como a medicina, a biotecnologia, a agronomia, etc. (BORÉM; SANTOS, 2001).

Neste projeto serão integrados os softwares: Emboss, Pipegen, AutoDock, Gramm, Modeller, no intuito de colaborar com os pesquisadores da área de genética, biologia molecular, biotecnologia e bioinformática.

Cada software possui funções variadas, sendo o EMBOSS um software capaz de integrar uma série de pacotes e ferramentas disponíveis para análise de sequências. Já o PIPEGEN uma ferramenta que facilita a transferência de dados permitindo ao usuário realizar análises de diferentes programas, ele combina esses dados e obtêm um resultado final.

O AutoDock bem como os outros softwares é direcionado para auxiliar a área da genética, possui ferramentas automatizadas para acoplamento, permitindo identificar as ligações de moléculas a um receptor em uma estrutura 3 D.

A outra ferramenta que será integrada é o GRAMM, utilizado para o acoplamento de proteínas, e para prever a estrutura de um conjunto de proteínas que necessita somente das coordenadas atômicas de duas moléculas, desprezando a necessidade de se conhecer informações sobre as suas ligações.

O Modeller é uma base de dados capaz de modelar estruturas de proteínas tridimensionais e a monta de acordo com a técnica de satisfação de restrições espaciais, o programa é usado para homologia ou para comparar modelos de estruturas de proteínas, o usuário fornece um alinhamento de uma sequência a ser modelado com estruturas relacionadas conhecidas e o Modeller automaticamente calcula um modelo com todos os átomos que não sejam hidrogênio.

Contudo podemos perceber a importância e a necessidade da integração da informática com as diversas áreas biológicas.

Os softwares foram instalados e testados em uma máquina virtual (software que virtualiza sistemas operacionais, ou seja, permite rodar outros sistemas dentro de um já instalado), como a plataforma foi desenvolvida voltada para a web foi realizada a virtualização do sistema operacional gratuito e voltado para servidores o Ubuntu Server, foi instalado no Ubuntu o programa Apache para que pudesse se iniciar a programação da plataforma para a web, a partir disso, iniciou-se o processo de programação visual da página, utilizando o Html juntamente com o framework bootstrap(responsável por aplicar um estilo artístico ao Html e tornar a página responsiva, que se adapta de acordo com o tamanho da tela).

Todo o código juntamente com o projeto foi desenvolvido no NetBeans IDE, um ambiente gratuito que fornece várias ferramentas para a integração com a web de desenvolvimento de código que abrange várias linguagens, inclusive a Php, que foi a linguagem utilizada no projeto. O NetBeans ainda fornece várias ferramentas para a integração com a web.

A partir da instalação do NetBeans foi realizada toda a programação estrutural das funcionalidades utilizando a linguagem Php, posteriormente integrada com a Html, juntando corpo(Html) e cérebro(Php) da plataforma.

Após a realização das junções foi necessário realizar a programação de códigos em Shell Script, uma linguagem de script utilizada para tornar mais simples

as execuções de tarefas repetidas, ou seja, para automatizar as execuções dos programas que foram integrados. Alguns programas necessitaram da instalação da extensão expect, uma ferramenta que automatiza processos que recebem comandos interativos, ou seja, que todos os dados não são passados em uma única linha de comado no terminal, que necessitasse acrescentar novas informações após a execução do programa.

Posteriormente integrou-se os scripts no código php, utilizando o comando shell_exec, que executa comandos em shell e retorna a saída em forma de string. Após todas as junções de código realizou-se a execução da plataforma localmente com o servidor Apache, após isso foram realizados testes de funcionalidade dos programas com exemplos já conhecidos.

6.RESULTADOS E DISCUSSÕES

Os dados gerados em laboratórios apresentam grande extensão, o que torna fundamental a utilização de softwares para a análise dos dados obtidos, a fim de evitar redundâncias e erros. Através da construção da plataforma web proposta, é possível realizar a análise de experimentos dos softwares apresentados em um único domínio (EMGPA – Emboss Modeller Gramm Pipegen Autodock).

Uma grande vantagem de se utilizar a EMGPA (figura 1) está na facilidade de desenvolvimento sem a necessidade de instalar qualquer programa.

Figura 1 - Página inicial da plataforma web.

Na plataforma web é possível verificar o local onde estão dispostos de maneira acessível os softwares Emboss, Modeller, Pipegen, Autodock e Gramm que foram integrados de modo a facilitar a pesquisa cientifica possibilitando o acesso simplificado e ágil onde com um clique é possível escolher o programa pretendido. (figura 2).

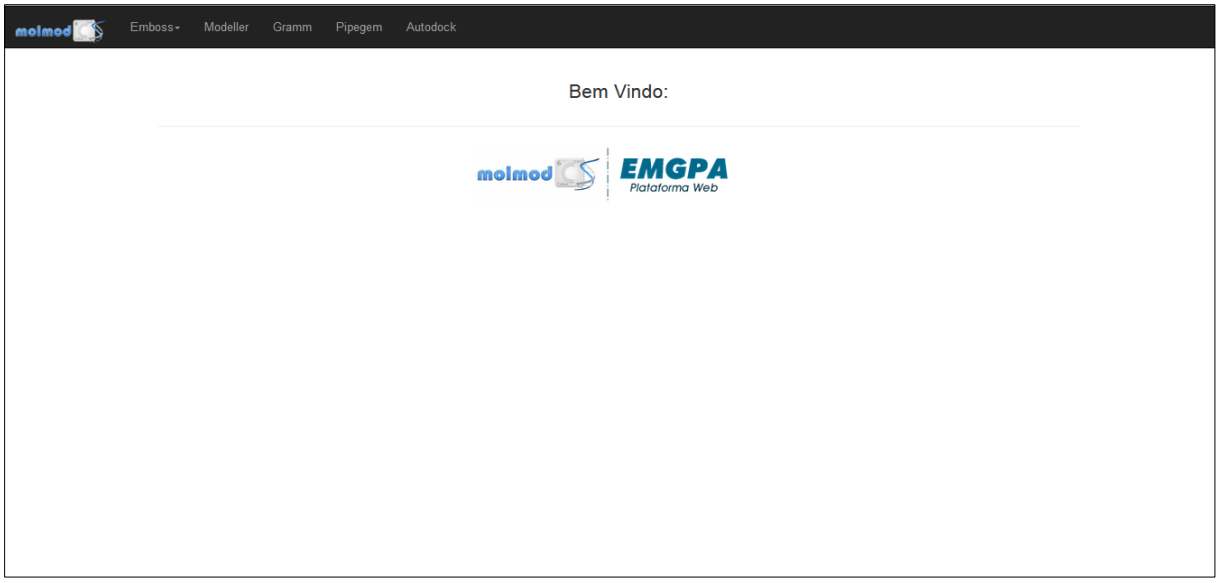

Figura 2 - Página de acesso aos programas

O Emboss é apresentado de maneira mais simples, como exemplo o Needle onde o pesquisador escolhe as duas sequências de entrada e define os “Gaps” se necessário ou os deixa como padrão, realizando assim o procedimento do Needle graficamente, numa plataforma onde não é necessária a instalação e execução em linha de comando (Figura 3).

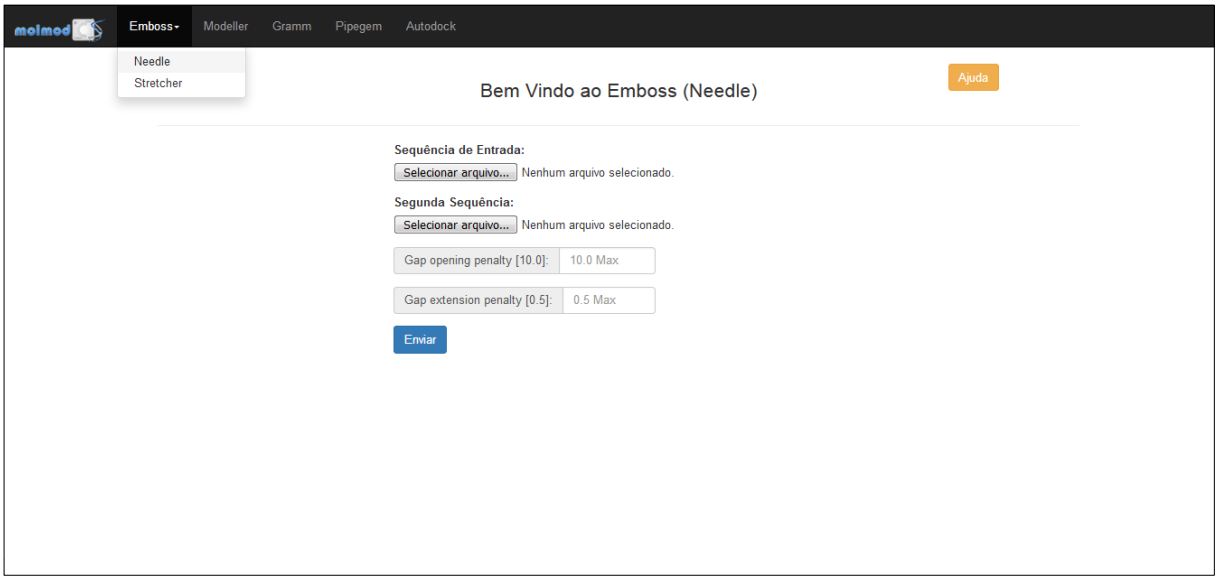

Figura 3 – Página Needle

Uma acessibilidade maior se dá ao clicar em ajuda (Figura 3) onde é apresentado a função do programa, sua descrição, a maneira de uso e um link para mais informações contendo a página do referido programa. (Figura 4)

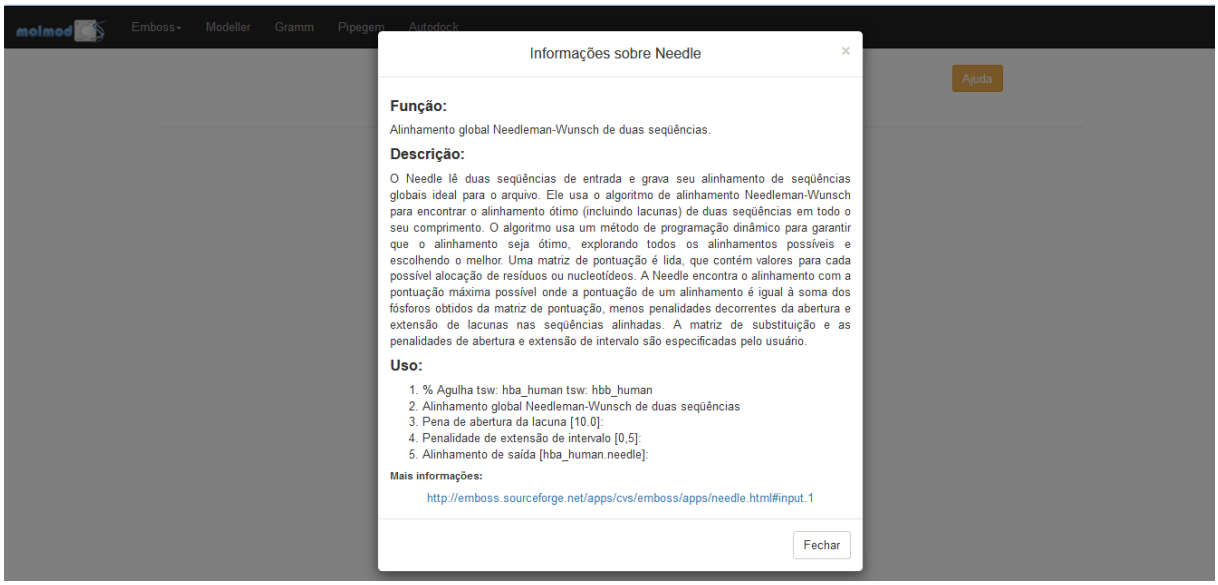

Figura 4 - Ajuda do Needle

Um recurso adicional que está presente na plataforma é a realização do download do resultado obtido, que poderá ser feito através da solicitação do usuário, dando um recurso a mais ao pesquisador que antes com o programa não possuía. (Figura 5)

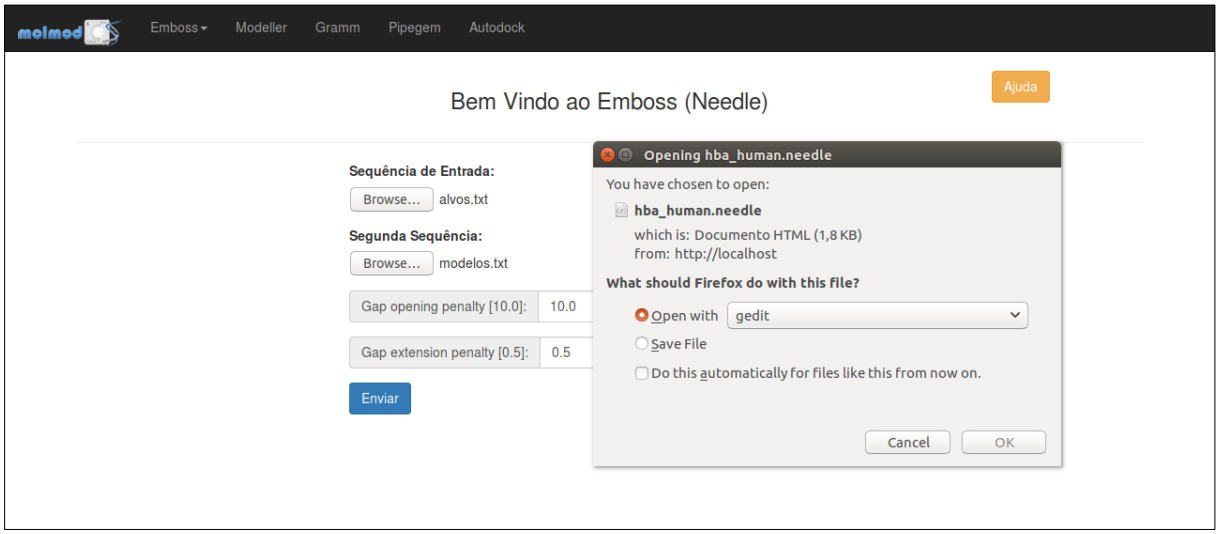

Figura 5 - Download resultado

A plataforma web EMGPA apresenta os programas de uma forma gráfica mais simples, onde não se necessita ter conhecimento sobre a instalação do software, que muitas vezes possui execuções complexas, o usuário apenas tem que possuir o entendimento da funcionalidade do programa, já que a plataforma apresenta uma navegabilidade fácil e intuitiva, de acesso rápido aos dados, possuindo recursos adicionais que facilitam a pesquisa do usuário.

7. CONCLUSÃO

A integração de softwares através de sistemas web simplifica a maneira com que os pesquisadores trabalham, por agrupar todos os softwares em um só local, facilita as pesquisas, pois a integração permite que um software da plataforma realize experimentos e os resultados obtidos já possam ser seguidamente utilizados em outro software da plataforma, não necessitando fazer uso de diferentes softwares dispostos em distintos locais, que muitas vezes possuem instalação, execução e processamento complexos dificultando o trabalho do pesquisador.

Através desse projeto, é evidente a necessidade da utilização da plataforma web de integração dos softwares Emboss, Modeller, Pipegen, Autodock e Gramm, pois a mesma permitiu um melhor gerenciamento das informações de pesquisas, otimizando o tempo e obtendo maior produtividade.

8. REFERÊNCIAS BIBLIOGRÁFICAS

AHRENS CH, BRUNNER E, QELI E, BASLER K, AEBERSOLD R. **Generating and navigating proteome maps using mass spectrometry.** Nat Rev Mol Cell Biol**.** 11(11):789-801, 2010.

AZEVEDO, JR. W.F.; MUELLER-DIECKMANN, H.J.; SCHULZE-GAHMEN U.; WORLAND, P.J.; SAUSVILLE, E.; KIM, S.H. **Structural basis for specificity and potency of a flavonoid inhibitor of human CDK2, a cell cycle kinase.** Proc Natl Acad Sci. USA 3:2735–2740. 1996.

BORÉM, A.; SANTOS, F.R. **Biotecnologia Simplificada**. Editora Suprema. Viçosa, MG. 2001.

BOYSEN, C.; SIMON, M.I.; HOOD, **L. Fluorescence-based sequencing directly from bacterial and P1-derived artificial chromosomes.** Biotechniques. V. 6, p. 978-82, 1997.

CORTHALS, G. L.; WASINGER, V. C.; HOCHSTRASSE, D. F.; SANCHEZ, J. C**. The dynamic range of protein expression: A challenge for proteomic research.** Electrophoresis, v. 21, p. 1104-1115, 2000.

DALL'OGLIO, Pablo. **PHP Programando com Orientação a Objetos 3ª Edição**. Novatec Editora, 2015.

DAVIES, K.; DHIMAN, N.; SMITH, D.I.; POLAND, G.A. **Decifrando o genoma: a corrida para desvendar o DNA humano.** Next-generation sequencing: a transformative tool for vaccinology. Expert Rev Vaccines. V.8, p. 963-7, 2001.

FLEISCHMANN, R. D.; ADAMS, M.D.; WHITE O.; CLAYTON,R.A.; KIRKNESS E.F.; KERLAVAGE, A.R.; BULT, C.J.; TOMB, J.F.; DOUGHERTY, B.A.; MERRICK, J.M**. Whole- genome random sequencing and assembly of Haemophilus influenzae** Rd. Science 269 (5223), 496- 512, 1995.

GALDOS, A. C. R. **Análise proteômica do saco vitelino de bovinos.** [Proteomic analisys of bovine yolk sac]. 2009. 111 f. Dissertação (Mestrado em Ciências) – Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, São Paulo, 2009.

GÖRG, A.; KLAUS, A.; LÜCK, C.; WEILAND, F**. Two-dimensional electrophoresis with immobilized pH gradients for proteome analysis: A laboratory manual.** 2010.

HAYNES, B.; CHEUNG, A.; BALAZINSKA, M.; **PipeGen: Data Pipe Generator for Hybrid Analytics.** Department of Computer Science & Engineering University of Washington. arXiv:1605.01664v2 [cs.DB] 15 May 2016.

HOSOUME, J.M. **Ação de anestésicos gerais em canais iônicos.** 2016. Dissertação (Mestrado em Biologia Molecular). Universidade de Brasília, Brasília, DF, 2016.

JENSEN, O.N. **Modification-specific proteomics: characterization of posttranslational modifications by mass spectrometry.** Curr Opin Chem Biol. V.8(1) p:33-41, 2004.

KLOSE, J. **Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A Novel Approach to Testing for Induced Point Mutations in Mammals.** Humangenetik, v. 26, p. 231-243, 1975.

MARQUI, AB.; VIDOTTO, A.; POLACHINI, G.M.; BELLATO, C.M.; CABRAL, H; LEOPOLDINO, A.M. **Solubilization of proteins from human lymph node tissue and twodimensional gel storage.** J Biochem Mol Biol. v.39(2), p. 216-22, 2006

MORRIS. Título: Auto Dock Disponível em:< **http://autodock.scripps.edu/**>. 07/10/2016.

NIEDERAUER, Juliano. PHP para quem conhece PHP. **São Paulo: Novatec**, 2004.

PANDEY, A.; MANN, M. **Proteomics to study genes and genomes.** Nature, Massachusetts, v. 405, n. 6788, p.837-846, 2000.

QUALTIERI, A.; PERA, M. L.; URSO, E.; BONO, F.; VALENTINO, P.; SCORNAIENCHI, M. C.; QUATTRONE, A. **Two-dimensional electrophoresis of peripheral nerve proteins: optimized sample preparation.** Journal of Neuroscience Methods, v. 159, p. 125-133, 2007.

RABILLOUD, T. **Two dimensional gel electrophoresis in proteomics: old, old fashioned, but still climbs up the mountains.** Proteomics, v. 2, p. 3-10, 2002.

RICE,P.; LONGDEN, I.; BLEASBY, A. ***EMBOSS:*** **The European Molecular Biology Open Software Suite**. Trends in Genetics 16, (6) pp276—277, 2000.

RIVEIROS, A.C.G.; PIZA, A.R.T.; RESENDE, L.C.; MARIA, D.A.; MIGLINO, M.A. Proteômica: **Novas fronteiras na pesquisa clínica enciclopédia biosfera**. Centro Científico Conhecer - Goiânia, v.6, n.11, p. 1, 2010.

ROCHA, T. L.; COSTA, P. H. A.; MAGALHÃES, J. C. C.; EVARISTO, R. G. S.;VASCONCELOS, E. A. R.; COUTINHO, M. V.; PAES, N. S.; SILVA, M. C. M.; GROSSI-DE-SÁ, M. F. **Eletroforese bidimensional e analise de proteomas**. Embrapa Recursos Genéticos e Biotecnologia, n. 136, p. 1- 12, 2005.

SALI, A.; POTTERTON, L.; YUAN, F.; VAN VLIJMEN H.; KARPLUS M. **Evaluation of comparative protein modeling by MODELLER**. Proteins, v. 23, n. 3, p. 318-26, Nov 1995.

SANGER, F.; NICKLEN, S.; COULSON, A.R. **DNA sequencing with chain-terminating inhibitors.** Proc Natl Acad Sci USA. v. 74(12), p. 5463-7, 1997.

SAVOIA, Hugo Rossetti. **XHTML e CSS+ PHP e MySQL**. IELD TEC, 2013.

SILVA, W.M.C**. Desenvolvimento de plataforma WEB para a análise e integração de dados proteômicos).** Consócio Setentrional Licenciatura em Biologia- UNB, 2011.

SILVEIRA, N.J.F.; UCHOA, H.B.; PEREIRA, J.H.; CANDURI, F.; BASSO, L.A.; PALMA, M.S.; JUNIOR, W.F.A. **Molecular models of protein targets from Mycobacterium tuberculosis.** Journal of Molecular Modeling, 2005.

SIMPSON, A.J.; REINACH, F.C.; ARRUDA, P.; ABREU, F.A.; ACENCIO, M.; ALVARENGA, R.; ALVES, L.M.; **The genome sequence of the plant pathogen Xylella fastidiosa.** Nature 406: 151- 157, 2000.

SIZOVA, D.; CHARBAUT, E.; DELALANDE, F.; POIRIER, F.; HIGH, A. A.; PARKER, F.; DORSSELAER, A. V.; DUCHESNE, M.; DIU-HERCEND, A. **Proteomics analysis of brain tissue from an Alzheimer´s disease mouse model by two-dimensional difference gel electrophoresis.** Neurobiology of Aging, v. 28, p. 357-370, 2007

SNUSTAD, P.; SIMMONS, M. J. **Fundamentos de Genética**. Ed. Guanabara Koogan S.A. Rio de Janeiro, RJ, 756p. 2001.

STEINDORFF, A.S. **Genômica estrutural e funcional de fungos do gênero *trichoderma.*** Universidade de Brasília. Programa de pós-graduação em biologia molecular. Brasília, 2016.

STULTS, J. T.; ARNOTT, **D. Proteomics**. Methods Enzymology, v. 402, p. 245–289, 2005.

TOVCHIGRECHKO A., VAKSER I.A.; **Servidor web público GRAMM-X para acoplamento de proteínas e proteínas.** *Nucleic Acids,* 2006.
Doi: 10.1093 / nar / gkl206

USAMI, M.; MITSUNAGA, K.; NAKAZAWA, K. **Two-dimensional electrophoresis of protein from cultured postimplantation rat embryos for development toxicity studies.** Toxicology in Vitro, v. 21, p. 521-526, 2007.

VALLEDOR, L.; JORRIN, J. **Back to the basics: maximizing the information obtained by quantitative two dimensional gel electrophoresis analyses by an appropriate experimental design and statistical analyses.** J Proteomics. V.74(1), p.1-18, 2010.

VENTER, J.C.; ADAMS, M.D.; MYERS, E.W.; LI, P.W.; MURAL, R.J.; SUTTON, G.G.; SMITH, H.O.; YANDELL, M. **The sequence of the human genome.** Science. 291: 1304-135, 2001.

VICKLUND, A. **Proteína-Proteína Docking e Protein-Ligand Docking.** Gama Global Correspondência Molecular, Vakser Lab. 2008-2017