A BIOINFORMÁTICA E O DESENVOLVIMENTO DE NOVAS ABORDAGENS COMPUTACIONAIS PARA SUPRIR O AUMENTO DE VOLUME DE DADOS DAS PESQUISAS

Evelyn Aparecida Gomes
Matheus Henrique Lopes Badoco[1]

Resumo

Este artigo tem por objetivo mostrar o aperfeiçoamento e o surgimento de novos avanços tecnológicos que vão desde *hardwares*, subáreas da bioinformática, até os *softwares* mais apropriados para catalogar, organizar e estruturar os conjuntos de dados que interagem com um banco de dados projetado para suprir o fluxo de volumes de análises coletadas por pesquisas voltadas para a área de biologia molecular. Este relacionamento tem ênfase nas informações projetadas, de modo a facilitar consultas, atualizações de pesquisas, gráficos, atendimento a necessidade de se manipular os dados genéticos e bioquímicos, resultando na construção de ferramentas, comparações feitas por algoritmos desenvolvidos para esta função e metodologias aplicadas no armazenamento dos dados e consequentemente, a tecnologia abordada na integração da biologia computacional. Assim, pretende-se, com base nestas circunstâncias, obter um crescimento e melhor desenvolvimento de técnicas e equipamentos sofisticados para alocar as pesquisas em campo. Para atingir os nossos objetivos, a metodologia empregada nesta pesquisa é de caráter bibliográfico e se baseia nos estudos de Shortliffe e Cimino (2006) e nos estudos de Rocha (2011), para falar de banco de dados em bioinformática; o Lesk (2008), para estudar os conceitos de bioinformática e partes biológicas e as pesquisas de Lorena e Carvalho (2003), para a utilização de técnicas inteligentes em bioinformática, e o manual do usuário da revista *Biotecnologia Ciência & Desenvolvimento*, para tratar sobre a bioinformática, sistemas operacionais, linguagens de programação e banco de dados; além disso, baseamo-nos na dissertação do mestrado de Miranda (2011), para obter informações sobre informática de bioimagens e, por fim, o artigo de Farias *et.al.* (2008).

Palavras-chave: Banco de Dados. Biologia Molecular. Dados Genéticos. Integração.

Abstract

This article has the objective to demonstrate the improvement and releases of new technological advances that goes from hardware, bioinformatics sub areas, to software to catalog, organize and structure lots of files that Interact with a data bank, projected to deal with the analysis collected from molecular biological researches. This relation had emphasis on the information projected, with the objective to make searches, researches updates, graphics, need to manipulate genetic and

[1] Alunos matriculados no segundo semestre do Curso de Sistemas de Informação do Uni-FACEF, Centro Universitário de Franca.

biochemical data easier, resulting in a construction of tools, comparisons made from algorithms designed for this use and methodologies applied on data storing and, consequently, the technology used in the biological computing integration. Under these circumstances, we intend to obtain an improving to develop better techniques and sophisticated equipment to allocate camp researches. To reach our objectives, the methodology in this article is based on Shortlife's and Cimino's book (2006), and on Rocha researches (2011), to talk about Data Bank in Bioinformatics, Lesk's book (2008), to study about concepts in bioinformatics and biological parts, in the studies of Lorena and Carvalho (2003), for intelligent technological uses in Bioinformatics, on the user manual from the magazine "Biotecnologia Ciência & Desenvolvimento", to learn about bioinformatics, operational systems, programming languages and Data Bank; besides, we also based on Miranda's article (2011), to obtain information about bio images in informatics and, at last, Farias's et.al. article (2008).

Keywords: Database. Molecular Biology. Genetic Data. Integration.

1 Introdução

No presente artigo, optamos por estudar a área de bioinformática, que estabelece modelos lógicos, matemáticos e estatísticos para decifrar os códigos genéticos. Esta área está em uma crescente expansão, devido aos avanços da medicina em busca de pesquisas para descobertas da cura de diversas doenças, que acarretam um vasto aumento nos fluxos de informações geradas que precisam ser comparadas e, consequentemente, armazenadas em um banco de dados para consultas e confrontamentos.

O objetivo desde artigo é realizar um estudo sobre o aperfeiçoamento e surgimento de novas técnicas computacionais que visam suprir o aumento de volumes de dados gerados nas pesquisas laboratoriais voltadas para a biologia. Estas informações são dispostas de modo a facilitar a consulta, a analisar imagens e gráficos, a manipular dados.

Este artigo foi produzido por meio de uma pesquisa de caráter bibliográfico, na qual utilizamos os estudos de Rocha (2011), Oikawa (2011), Felofiloff (2015) e Lifschitz (2006), para abordar os conceitos de banco de dados na área genômica, além dos conceitos citados nos estudos de Lorena (2003), para tratar dos conceitos de DNA, RNA, proteínas e algumas aplicações e técnicas em bioinformática.

Baseamo-nos nos livros de Lesk (2008), para estudar os conceitos de bioinformática e as partes biológicas como, por exemplo, a organização e a evolução

dos genomas. O Manual do Usuário da *Revista de Bioinformática* para tratar sobre os sistemas operacionais, as linguagens de programação utilizadas por um profissional da área, abordagem de banco de dados com recursos SGBD e banco de dados públicos.

Por fim, a dissertação de Miranda (2011), para alavancar os estudos sobre bioimagens. Utilizamos as propostas de Kvilekval (2009), para tratar sobre a ingestão de imagens e metadados, estudo da ferramenta *Bisque,* que disponibiliza uma plataforma *web* para análises e serviços baseados nas imagens obtidas nas pesquisas.

Este trabalho se inicia com uma abordagem dos conceitos da área, como surgiu a bioinformática e a necessidade de desenvolver programas que supram o aumento de volumes de dados e consigam reconhecer as sequências de genes, análises tridimensionais para identificar e ou organizar e relacionar informações biológicas a partir de dados já armazenados em um banco de dados que permita uma rápida resposta. Em seguida, será tratada a importância de manter os registros dos dados coletados, os pontos de deficiência do sistema de gerenciamento de banco de dados, como são realizadas as coletas de informações e a organização dos algoritmos.

E, na seção quatro, trataremos sobre a subárea de bioinformática que analisa as imagens tridimensionais, por meio de uso de ferramentas avançadas que possuem *softwares* com conceitos inovados, que permite como solução o uso de técnicas que facilitam as transferências e processamento rápido.

2 Conceito de Bioinformática

Conforme nos diz Lesk (2008), buscando tratar dados biológicos brutos, a bioinformática ou biocomputação surgiu em meados da década de 1980 quando se consolida como uma nova área do conhecimento, devido ao aumento da necessidade de desenvolver programas computacionais que permitam reconhecer sequências de genes, configuração tridimensional de proteínas, identificar inibidores de enzimas, organizar e relacionar informação biológica, agrupar proteínas homólogas, estabelecer árvores filogenéticas, analisar experimentos de expressão gênica, dentre outros.

A bioinformática tem seu maior foco nos estudos de citômica, genônima, genômica funcional, proteômica e metabolômica de seres vivos e dos seres humanos, ou seja, todas com o sufixo “ômica” visam à identificação funcional ou estrutural nas moléculas que possuem estruturas como RNA, DNA ou proteínas.

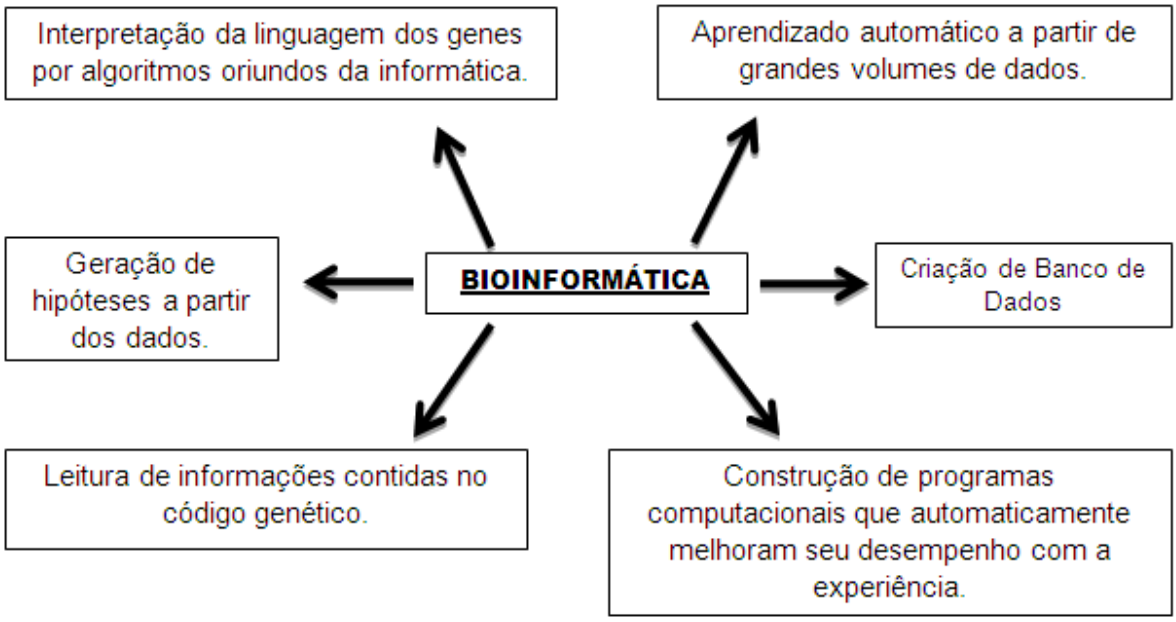
Um gene está localizado na região da molécula do DNA (ADN, ácido desoxirribonucleico, ou DNA, deoxyribonucleic acid) que contém uma instrução genética codificada através da base nitrogenada, formação de molécula de RNA (ARN, ácido ribonucleico, ou RNA, ribonucleic acid) responsável pela síntese de proteínas que controlam as estruturas, funções metabólicas e todo o organismo.

A desfragmentação do DNA é composta por quatro partes denominadas pelas letras A, C, T, G, que em conjuntos formam os sequenciadores. Os sequenciadores dos genomas de diferentes organismos passaram a ser produzidos em grandes quantidades pelas pesquisas (Lorena, 2003). Devido a este aumento, houve a necessidade de armazenamento em um banco de dados e de comparações realizadas por algoritmos, pois, durante um processo de pesquisa, uma nova sequência de genes é comparada com a que está armazenada no banco de dados da mesma função, permitindo uma rápida resposta, devido ao grande fluxo de informações. Rubino (2012) nos diz que essas informações dos genes podem ser encontradas em bancos de dados *online*, em que os pesquisadores colocam suas descobertas para o público, mostrando onde o gene atua, qual proteína ele ajuda a produzir, em qual função está presente, localização celular, processo metabólico, via metabólica e localização no cromossomo, utilizando conhecimentos através de métodos experimentais como técnicas *in vitro*, *in silico* e ou *in vivo*.

Atualmente, a qualidade de um banco de dados depende não apenas da informação que contém, mas de suas conectividades com outras fontes de informação que os usuários podem acessar e extrair informações, ou submeter materiais para processamento pela equipe dos bancos, mas não podem adicionar ou alterar entradas dos bancos de dados diretamente. Como efeitos ambientais exercem controle sobre os eventos genéticos, o fluxo e o controle da informação de um organismo, nos dias de hoje, são acessíveis pela bioinformática, portanto estas informações entram no domínio da bioinformática quando um cientista deposita seus resultados experimentais em arquivos de banco de dados apropriados, conforme se lê em Lesk (2008).

Abaixo apresentamos um fluxograma que explica como são formados os conceitos e interpretação dos dados que originaram a bioinformática.

Figura 1 – Associação de conhecimentos entre biologia molecular e informática, que culminou com o advento da bioinformática.



Fonte: ARAÚJO, 2008, p. 04.

3 A importância do Banco de Dados em Bioinformática

A importância de um banco de dados é manter o registro de tudo que foi realizado, e, em bioinformática, isso não é diferente.

> A bioinformática torna-se essencial para a construção de bases de dados contendo informações sobre os genes e proteínas dos organismos vivos, para a descoberta de novos genes, e de novos medicamentos, pois é através da bioinformática que novas técnicas para o mapeamento e armazenamento das informações extraídas dos genes vêm sendo estudadas e estruturadas. (ROCHA, 2011, p.04).

Um dos seus grandes problemas é a falta de um sistema de Gerenciador de Bancos de Dados (também conhecido como SGBD) específico para a área, o que impossibilita beneficiar-se de mecanismos mais eficientes de armazenamento e gerenciamento inteligente de memória.

3.1.1 Coleta e análise de dados

Para a coleta de dados, são utilizadas técnicas avançadas de computação, para que possa ser criado um bom banco de dados mais fácil para ser analisado. Já no caso da análise de dados, é necessário transformar técnicas experimentais, dentre elas: nanotecnologia e microscopia modernas, em conhecimento.

As técnicas para gerenciamento de dados são fundamentais para o desenvolvimento de aplicações biológicas, pois oferecem métodos adequados para gerenciar, projetar e acessar os dados. Vale observar que as técnicas de gerenciamento tradicionais podem não ser adequadas para gerenciar dados biológicos.

3.1.2 Composição, definição e criação de um banco de dados

Um sistema de banco de dados inclui todos os componentes que são envolvidos na organização dos recursos, incluindo os próprios dados, o *software* do SGBD (Gerenciador de Bancos de Dados), o *hardware* do sistema, a mídia de armazenamento e os aplicativos que acessam e atualizam os dados.

Um banco de dados pode ser definido como uma coleção de dados logicamente relacionados, projetado para atender às necessidades de informação de diversos usuários, armazenando todas as informações para serem analisadas ou observadas futuramente. O SGBD é uma coleção de componentes de *software* para gerenciar, criar e consultar informações, contendo diversos bancos de dados.

O processo de criação de um banco de dados é iniciado com a análise das informações que ele deve conter: uma definição dos tipos de dados, relação entre as informações e restrições. O objetivo de projetar um banco de dados é de especificar a estrutura lógica e física de um ou mais banco de dados, acomodando a informação necessária para os usuários em um conjunto definido de aplicações como nos mostra Oikawa (2011).

Os algoritmos usados na bioinformática possuem o principal objetivo de comparar duas sequências diferentes para identificar a similaridade entre elas (Sequência de DNA, por exemplo). A ideia central das técnicas desenvolvidas para

análise é de minimizar as diferenças entre elas, ou seja, buscar um alinhamento perfeito.

Um grande exemplo do uso desses algoritmos é no caso do DNA: analisar duas sequências diferentes de DNA, e identificar as proteínas presentes em cada ambas.

> Para a recuperação de sequências semelhantes, precisamos medir a semelhança da sequência sonda com cada sequência do banco de dados. É possível fazer muito melhor do que a solução simples de checar cada par de posições em cada justaposição possível, um método que, mesmo sem permitir a inserção de lacunas, exigiria um tempo proporcional ao produto do número de caracteres na sequência sonda pelo número de caracteres no banco de dados. Uma especialização da ciência da computação, conhecida vulgarmente como "*stringology",* concentra-se no desenvolvimento de métodos eficientes para este tipo de problema, analisando seus desempenhos efetivos (LESK, 2005, p.360).

Segundo Feofiloff (2010), "A análise de algoritmos estudam a correção e o desempenho de algoritmos. Em outras palavras, a análise de algoritmos procura respostas para perguntas do seguinte tipo: Este algoritmo resolve o meu problema? Quanto tempo o algoritmo consome para processar uma entrada de tamanho n?

De modo geral, os algoritmos são utilizados apenas para comparação entre duas sequências, diferentemente dos *softwares,* que são utilizados para análise, separadamente. A vantagem da utilização desses algoritmos é o fato de conseguir realizar uma análise precisa sobre duas sequências.

3.1.3 Banco de Dados Primário e Secundário

Em estudo recente, Seibel (2006, p. 11) afirma que “os bancos de dados aplicados á biologia molecular poder ser classificados de acordo com as informações biológicas que armazenam”.

As informações biológicas podem estar presentes nas sequências de nucleotídeos ou de proteínas; podem ser proteínas com informações sobre suas funções; taxonomia que classifica os organismos vivos e bibliografia que contem os artigos, jornais e periódicos.

> Os bancos de dados primários apresentam resultados de dados experimentais que são publicados com alguma interpretação, mais não há uma análise cuidadosa desses dados. Esse é o caso, por exemplo, do GenBank, EMBL e PDB (Protein Data Bank). Já os secundários são aqueles onde há uma compilação e interpretação dos dados de entrada por um ou mais grupos de cientistas, de forma que podem ser obtidos dados mais

> representativos e interessantes. Esses são os bancos de dados curados, como o COG, SWISS-PROT e o TrEMBL. (PROSDOCIMI, 2007, p. 45).

Abaixo, o fluxograma mostra todo o ciclo do processo realizado desde a pesquisa até a publicação das informações obtidas pelas comunidades.

Figura 2 - Fluxograma de Informações e alimentação do banco de dados.

Fonte: RUBINO, 2013, p. 07.

4 BioImagens

Os grandes avanços de pesquisas genômicas acarretam no avanço tecnológico, pois, as funções das proteínas dependem da adoção de uma estrutura tridimensional (3D) do seu estado nativo; a estrutura nativa de uma enzima pode apresentar uma cavidade na sua superfície, que se liga a uma pequena molécula e a coloca próxima de seus resíduos catalíticos, em que os mecanismos dependem da ligação de proteínas ou do DNA, focalizando na análise de dados relacionados a estes processos. Para sua melhor identificação, são realizadas as triagens virtuais.

4.1.1 Conceito e funcionalidades de Bioimagens

A bioimagem, nos conceitos de Kvilekval (2009) é uma subárea da bioinformática utilizada para análises tridimensionais, conforme citado acima. Contudo, existe uma ferramenta avançada neste seguimento, conhecida como *Bisque*, que fornece um sistema de análise baseadas em imagens. Divididos em dois métodos podemos observar que os métodos internos são executados através da *web*, facilitando o uso durante a transferência para os servidores do *Bisque* e nos métodos de análises externos são utilizadas ferramentas para acessar remotamente dados dentro do sistema onde podem ser utilizados hardwares e também são úteis para a prototipagem rápida.

O principal diferencial do *Bisque* é o fornecimento de recursos *online* para gestão e análise de imagens biológicas em 5D e o gerenciamento de coleção de imagens, todos facilitam os fluxos de pesquisas biológicas conforme nos mostra Fedorov (2009).

Este sistema é projetado de forma que permite o suporte e compartilhamento de imagens a partir de um servidor de imagens (IS) para armazenar, e ou recuperar arquivos, a partir de servidores de dados (DS) que tem função de processamento, redimensionamento das imagens. As imagens são requisitadas e colocadas em *pipelines* (técnica utilizada para acelerar o processamento e velocidade da CPU), elas são armazenadas em *caches* para evitar a sobrecarga de transporte excessivo na rede. O sistema também produz *plug-ins* para interação das imagens que são importadas e exportadas para os servidores.

Os principais serviços incluem armazenamento de imagens e gestão, gerenciamento de metadados e consulta, execução de análise e apresentação do cliente. (KVILEKVAL, 2009)

4.1.2 ITK – *Toolkit Insight*

O ITK é utilizado pela medicina para análise de imagem, é um sistema *open source*, que fornece aos desenvolvedores uma gama de conjuntos de ferramentas de software para análise de imagem. O *software* emprega algoritmos para registrar e segmentar dados multidimensionais.

Ele tem total integração com o ambiente *Bisque* a fim de construir módulos específicos. Definido a partir de pipelines, o usuário cria um ambiente de processamento e desenvolvimento diferenciados. (KVILEKVAL, 2009).

4.1.3 BioView 3D

O *BioView* 3D é um outro sistema que permite ao usuário a visualização rápida de imagens em 3D. É possível fazer a importação juntamente com o software *Bisque*, em que, a partir da URL informada é carregado um sub arquivo que permite explorar os dados publicados fazendo *download*, análises e navegando dentro dos dados. (KVILEKVAL, 2009).

A exploração desses dados publicados é dividida entre módulos que rastreiam a ponta da raiz da imagem em uma sequência de tempo; o outro módulo classifica as células com base na expressão de helicoidais emparelhados (PHF) dos filamentos em imagens de fluorescência; o *CellProfiler* que também é um *software* livre permite, dentro do *BioView* 3D, que os biólogos meçam quantitativamente fenótipos de milhares de imagens automaticamente; existe um módulo que permite anotações gráficas; outro que renomeia tipos, nomes ou valores de anotações; tem o módulo que permite ao usuário encontrar uma segmentação em uma imagem (segmentação de imagens é normalmente usada para localizar objetos e limites em imagens), dentre outros.

Conclusão

Neste artigo o objetivo foi mostrar o aperfeiçoamento e surgimento de novas técnicas computacionais com intuito de suprir a grande demanda de pesquisas biológicas, diante disso acreditamos ter cumprido com esta proposta uma vez que descrevemos, pesquisamos, estudamos e descobrimos esta nova área que está em grande expansão tanto para área de biologia quanto para a área de tecnologia que traz.

A bioinformática teve seu surgimento a partir dos avanços na área de biologia molecular, especialmente com os estudos dos sufixos “ômicas” como a genômica, proteômica, metabolômica, dentre outras, que geraram um grande volume de informações obtidas pela fragmentação do DNA (ácido desoxirribonucleico). Partindo destas informações, a bioinformática permite o compartilhamento dos dados obtidos com a ajuda da internet por processadores mais rápidos e eficientes permitindo a otimização nos processos e uma interface relacionada com os dados já existentes que conseguem ser comparadas permitindo uma rápida resposta.

No entanto, com o avanço das pesquisas, a produção de dados aumentou e, consequentemente, os dados armazenados precisaram ser cada vez mais manuseados dando origem ao uso do Sistema de Gerenciamento de Banco de Dados (SGBD) para gerenciar o grande volume de dados. Os SGBD são *softwares* que permite o armazenamento e acesso eficientes aos dados.

O ápice do desenvolvimento de ferramentas que permitem uma inovação para a área é a interface gráfica apropriada para a visualização de imagens tridimensionais. Disposta em *softwares* que analisam *online* facilitam a prototipagem e um melhor suporte no compartilhamento das imagens.

Referências

ARAUJO, Nilberto Dias de et al. A Era da Bioinformática: *Seu potencial e suas implicações para as ciências da saúde*. Paraná, p. 01-06, Jan/Dez. 2008.

ESPINDOLA, Foued Salmen et al. Recursos de Bioinformática aplicados às ciências ômicas como genômica, transcriptômica, proteômica, interatômica e metabolômica, Uberlândia, v. 26, n. 3, p. 463-477, Jun. 2010.

FELOFILOFF, Paulo. Banco de Dados em Bioinformática, São Paulo. Disponível em: <http://www.ime.usp.br/~pf/analise_de_algoritmos/aulas/guloso.html>. Acesso em: 24 out. 2015.

GUIDO. Rafael V. C. et al. Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas, *Estudos Avançados*, São Paulo, v. 24, n. 70, p. 1-13, Out. 2010.

KVILEKVAL, Kristian et al. *Bisque*: a platform for bioimage analysis and management, USA, v. 26, n. 4, p. 544-552, Dez. 2009.

LESK, Arthur M. Introdução à Bioinformática. Tradução de Ardala Elisa Breda Andrade et. al. 2. ed. Porto Alegre: Artmed, 2008. 3

LIFSCHITZ, Sergio. Algumas pesquisas em Banco de Dados e Bioinformatica. 2006. Disponível em: <http://www.ebah.com.br/content/ABAAABGhsAD/algumas-pesquisas-bancos-dados-bioinformatica-sergio-lifschitz>. Acesso em 24 out. 2015.

LORENA, Ana Carolina et al. Utilização de Técnicas Inteligentes em Bioinformática. 2003, Universidade de São Paulo, São Carlos, 2003.

MIRANDA, Gisele Helena Barboni. Métodos para Processamento e Análise Computacional de Imagens Histopatológicas visando apoiar o Diagnóstico de Câncer de Colo do Útero. 2011, (Dissertação de Mestrado para obter grau de mestre em Ciências). Universidade de São Paulo, Ribeirão Preto, 2011.

OIKAWA, Márcio K.. Banco de Dados em Bioinformática. 2011. Disponível em: < http://www.ime.usp.br/posbioinfo/cv2011/marciooikawa.pdf>. Acesso em 24 out. 2015.

PROSDOCIMI, Francisco. Introdução a Bioinformática. 2007.Disponível em: < http://www2.bioqmed.ufrj.br/prosdocimi/FProsdocimi07_CursoBioinfo.pdf>. Acesso em 10 out. 2015.

ROCHA, Cícero Pinho. Banco de Dados em Bioinformática. 2011 (Artigo para obtenção de título de Especialista em Análises de Sistemas). UESPI, Parnaíba, 2011.

RUBINO, Gabriel; LOPES, Fabrício Martins. Integração de dados biológicos e redes gênicas: *um estudo de caso em arabidopsis thaliana.* 2012. Universidade Tecnológica do Paraná, Paraná, 2012.

SEIBEL, Luiz Fernando Bessa et. al. Banco de Dados Genoma. 2006. Disponível em: < http://www.uniriotec.br/~seibel/Tutorial%20SBBD2000%20FinalRev.pdf>. Acesso em 10 out. 2015.