UNIVERSIDADE FEDERAL DO PARANÁ

ALEXANDRE QUADROS LEJAMBRE

SILA EUKARYOTIC - FERRAMENTA PARA ANOTAÇÃO AUTOMÁTICA DE GENES EUCARIOTOS

CURITIBA

2016

ALEXANDRE QUADROS LEJAMBRE

SILA EUKARYOTIC – FERRAMENTA PARA ANOTAÇÃO AUTOMÁTICA DE GENES EUCARIOTOS

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, área de concentração Bioinformática.

Orientador: Prof. Dr. Roberto Tadeu Raittz

Coorientador: Prof. Dr. Vinicius Weiss

CURITIBA

2016

L534 LEJAMBRE, Alexandre Quadros
Sila Eukaryotic - Ferramenta para anotação automática de genes eucariotos/Alexandre Quadros Lejambre; orientador, Roberto Tadeu Raittz; coorientador, Vinicius Weiss. -- Curitiba, 2016.
59p.: il., color., tabs.

Dissertação (Mestrado) - Universidade Federal do Paraná. Setor de Educação Profissional e Tecnológica. Programa de Pós-Graduação em Bioinformática.

Inclui referências.

1. Genomas. 2. Genoma Eucariotos. 3. *Sila Eukaryiotic*. 4. Bioinformática. I. Raittz, Roberto Tadeu. II. Weiss, Vinicius III. Universidade Federal do Paraná. Programa de Pós-Graduação em Bioinformática. IV. Título.

CDD 574.0285

TERMO DE APROVAÇÃO

ALEXANDRE QUADROS LEJAMBRE

"SILA EUKARYOTIC - FERRAMENTA PARA ANOTAÇÃO AUTOMÁTICA DE GENES EUCARIOTOS"

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Dr. Roberto Tadeu Raittz
Universidade Federal do Paraná - UFPR

Dr. Rodrigo Luis Alves Cardoso
Bolsista PNPD-CAPES/Programa de Pós-graduação em Bioinformática – Universidade Federal do Paraná

Drª. Jeroniza Nunes Marchaukoski
Universidade Federal do Paraná - UFPR

Curitiba, 30 de setembro de 2016

Para a Humanidade.

AGRADECIMENTOS

Agradeço à minha família, em especial o meu avô Francisco Tomaz de Quadros pela amizade, ajuda e apoio; à minha avó (*in memorian*) Marilene de Mattos Quadros e aos meus pais Lili Marlene de Quadros Lejambre e D'Artagnan Lejambre pela criação, educação e incentivo.

Também agradeço ao meu orientador, Prof. Dr. Roberto Tadeu Raittz, pela oportunidade, paciência e apoio em todos os momentos. Ao meu coorientador, Dr. Vinicius Weiss, pelas dicas e ajudas durante o projeto. Deixo também meus agradecimentos aos Professores Dr. Dieval Guizelini, Dra. Jeroniza Marchaukoski e Dra. Rafaela Mantovani Fontana pelas dicas, sugestões e ensinamentos que me ajudaram neste projeto.

Aos meus amigos MSc. Bruno L. Nichio e MSc. Roxana Beatriz R. Chaves pela companhia, paciência e apoio durante esses anos de curso. Ao Programa de Pós-Graduação em Bioinformática-UFPR, CAPES, todos os meus colegas e professores do curso, bem como à Suzana Azevedo por toda atenção e disponibilidade em ajudar.

Agradeço também à mãe da minha filha, Larissa M. Lourenço, pela paciência e apoio, e finalmente à minha filha, Hellen Moreira Lejambre, que me mostrou como obter forças e persistência.

Muito obrigado a todos!

RESUMO

A marcação e anotação de genomas são processos essenciais e fundamentais presentes na Bioinformática, necessários para a análise de sequências de DNA. Geralmente essas atividades exigem muito tempo de processamento e alto custo computacional, encarecendo o processo nos muitos projetos distintos ao qual estão presentes. Tais atividades consistem basicamente em comparar sequências de DNA com grandes bancos de dados, este contendo entre milhares e milhões de registros. Considerando a necessidade de melhoramento dessas atividades fundamentais nas pesquisas que envolvem a Biologia Molecular, consequentemente a Bioinformática, este trabalho apresenta um programa de computador, chamado Sila Eukariotic, implementado para a anotação automática de sequências de DNA provenientes de organismos eucarióticos, utilizando busca e comparação de similaridades entre sequências de proteínas. O Sila Eukariotic utiliza ferramenta para predição e marcação de genes (GeneMark-ES) em conjunto com uma ferramenta de buscas de sequências de proteínas em banco de dados (RAFTS3), sendo a predição opcional ao usuário, cabendo também a opção pelo banco de dados de comparação a ser utilizado (NR, SwissProt e outros em formato multi-fasta). Os resultados evidenciaram melhoras na anotação de genes eucarióticos com baixíssimo custo e tempo de processamento comparado a outros buscadores disponíveis. O pacote está disponível para download e poderá ser executado em computadores com relativamente poucos recursos de hardware.

Palavras-chave: Anotação automática de genomas eucarióticos, Comparação de sequências, *Alignment-free*.

ABSTRACT

The genome annotations are important and essential processes present on Bioinformatics, necessary to analyse DNA's sequences. Usually these activities need a lot of processing time and have high computational costs, making the projects, which are present, remain expensives. These activities consisti, basically, in comparing DNA's sequences against big data banks, that contain between thousands to milions of records. Seeing the need to improve these activities, this work presents a computer's system, called Sila Eukaryotic, implemented to make the automatic annotation for sequences of eukaryotics organisms, using search and comparison of similarity between proteins sequences. Sila Eukaryotics uses a tool for gene prediction (GeneMark-ES) together with a search proteins tool (RAFTS3) at a biological data bank. The prediction is optional, as the data bank to be used (NR, SwissProt e others on fasta file format). The results showed improvements in annotation, quickly and low computational costs. This package is avaliable to download and can be executed in computers with few hardware resources.

Key-words: Automated eukaryotic genome annotation, Sequence comparison, Alignment-free.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

BCOM	Matriz binária de co-ocorrência de aminoácidos
BLAST	Basic Local Alignment Tool
BLAT	BLAST-Like Alignment Tool
C	Linguagem de programação de baixo nível
C++	Linguagem de programação descendente da linguagem C
CDS	Coding DNA Sequence
CEGMA	Core Eukaryotic Genes Mapping Approach
COG	Clusters of Orthologous Groups
DBSTRUCT	Estrutura do banco de índices do RAFTS3
DDBJ	DNA Databank of Japan
DNA	Deoxyribonucleic Acid (Ácido Desoxirribonucleico)
DOGMA	Dual Organellar Genome Annotator
DP	Dinamic Programming
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
EP	Estatística Probabilística
EST	Expressed Sequence Tag
FAS	Formato FASTA
FLN	Functional Linkage Network
GBK	Formato GenBank
GHMM	Generalized Hidden Markov Model
Ghz	Giga-hertz
GOA	GeneOntology Atributes
GPHMM	Generalized pair Hidden Markov Model
HMM	Hidden Markov Model
HSM-SVM	Hidden semi-Markov Suport Vector Machines
IMM	Interpoled Markov Model
INSDC	International Nucleotide Sequence Database Collaboration
KAAS	KEGG Automatic Annotation Server
KEGG	Kyoto Encyclopedia of Genes and Genomes
KOG	EuKaryotic Orthologous Groups (Específica versão do COG)
LOG	Na informática, é o registro de eventos computacionais
MCR	Matlab Compiler Runtime
MDD	Maximal Dependence Decomposition
MM	Markov Model
mRNA	RNA mensageiro
NBCI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NIH	National Institutes Health
NLM	National Library Medicine
NN	Neural Network
NR	Non-Redundant
ORF	Open Reading Frame
QDA	Quadratic Discriminant Analysis
RAFTS3	Rapid Alignment-Free Tool for Sequences Similarity Search
RAM	Random Access Memory
RNA	Ribonucleic Acid (Ácido Ribonucleico)
rRNA	RNA ribossomal
SNAP	(OH-SNAP) Optimized Hybrid Scale Neural Analog Predictor
tRNA	RNA transportador

SUMÁRIO

1 INTRODUÇÃO

A fim de conhecermos, entendermos e catalogarmos os genomas de organismos, sequências de DNA que possuem milhões ou bilhões de pares de base, que guardam as informações herdáveis e necessárias para a realização dos processos biológicos e sobrevivência do organismo, é necessário a análise das sequências desses genomas e suas proteínas resultantes das regiões codificantes. Tal análise envolve grande quantidade de dados que consequentemente resultam em muitos processos computacionais, exigindo altos recursos de hardware de um computador, além de que essas informações devem ser amazenadas e acessadas facilmente (BAXEVANIS; OUELLETTE, 2001).

Os bancos de dados de sequências, assim como a Bioinformática, iniciaram suas histórias no início da década de 60 com uma coleção de sequências protéicas, fruto das pesquisas da química Margaret Oakley Dayhoff, também conhecidas como Atlas de Sequências de Proteínas e Estruturas (BAXEVANIS; OUELLETTE, 2001). Conforme as tecnologias de sequenciamento foram desenvolvidas, houve um aumento exponencial de dados biológicos obtidos e armazenados em bancos de dados, cujo tornaram as análises manuais muito morosas, sendo transferidas então para o computador, surgindo assim o desenvolvimento da ciência denominada Bioinformática (PROSDOCIMI, 2007).

Então, considerando a grande quantidade de dados que necessitam de computadores robustos, propomos como objetivo geral deste trabalho o desenvolvimento de uma nova ferramenta integrada que faça a análise automática de genomas eucarióticos. Essa análise deverá ser realizada de forma rápida, sem necessitar de grande capacidade computacional, e que resulte na anotação das regiões codificantes, ou seja, marcando e identificando os genes existentes nas sequências de DNA. Além da anotação a ferramenta também tem como propósito a correção de regiões preditas a fim de melhorar a anotação final.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 BIOLOGIA MOLECULAR

Para melhor entendimento deste trabalho serão explicados alguns conceitos básicos sobre DNA, RNA, proteínas e o Dogma Central da Biologia Molecular. Essas explicações se limitarão ao foco da anotação de genomas.

2.1.1 DNA e RNA

O DNA (do inglês *deoxyribonucleic acid*) ou ácido desoxirribonucleico, é uma molécula que contém longas cadeias de sequências de nucleotídeos, necessárias para a produção de produtos orgânicos diversos que realizam os processos bioquímicos indispensáveis para a manutenção e sobrevivência de um organismo. Foi em 1953 que o DNA foi descrito pelos cientistas James D. Watson e Francis Crick, e sua estrutura foi demonstrada evidenciando sua disposição em forma de dupla hélice voltada para a direita, com cadeias sequenciais antiparalelas de nucleotídeos.

Os nucleotídeos são compostos por uma base nitrogenada, uma pentose e um fosfato. No DNA são encontrados quatro tipos de nucleotídeos: Adenina (A), Citosina (C), Guanina (G) e Timina (T). A ordem em que esses nucleotídeos estão encadeados determina as informações de sequências de outras moléculas orgânicas, produtos biológicos que atuam nos processos bioquímicos de um organismo. Esses produtos biológicos podem ser proteínas ou RNAs (Ácido Ribonucleico), e suas contruções originam-se de regiões específicas do DNA, chamadas de genes. A FIGURA 1 mostra a estrutura química das bases nitrogenadas que compõem o DNA e o RNA.

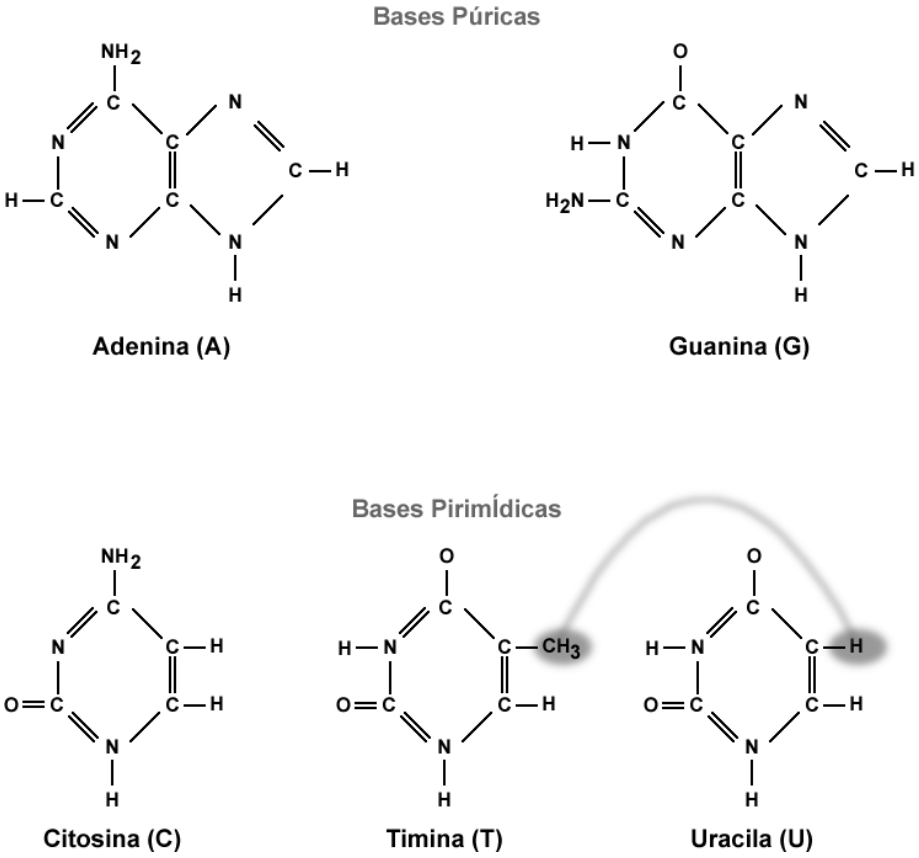

FIGURA 1 – ESTRUTURA QUÍMICA E CLASSIFICAÇÃO DAS BASES NITROGENADAS
FONTE: O Autor.

O RNA tem estrutura parecida com a do DNA, sendo diferentes no RNA o componente pentose e a ausência da base Timina, substituída pela base Uracila (U). RNAs podem ser de diferentes tipos e terem distintas funções (principais: mensageiro, transportador e ribossômico). O RNA mensageiro (mRNA) transporta informações até uma organela responsável pela síntese de proteínas chamada ribossomo, tendo este como seu principal componente o RNA ribossômico (rRNA). Nos organismos eucarióticos o mRNA precisa passar por processos de amadurecimento antes de interagir com o ribossomo, ou seja, o RNA não maduro sofre processos de *splicing* (cortes) que dividirão a molécula em regiões chamadas de éxons e íntrons. Os éxons são as partes da sequência que formarão o mRNA, já o íntrons possuem outras funções e não participam da sequência que determinará uma proteína. O RNA transportador (tRNA) é responsável pelo transporte de moléculas de aminoácidos até o ribossomo, esse transporte possibilita a construção de cadeias de aminoácidos (peptídeos) que formarão as estruturas de uma proteína (NELSON; COX, 2006).

2.1.2 Proteínas

Proteínas apresentam quatro níveis estruturais: a primária corresponde à própria sequência de aminoácidos; a secundária está relacionada ao arranjo espacial entre os aminoácidos, por exemplo, alfa-hélice e folha beta; na estrutura terciária está a conformação da proteína inteira; e a quartenária é a conformação de duas ou mais cadeias polipeptídicas. As proteínas são macromoléculas que estão presentes em todos os processos biológicos, desempenhando diversos papéis nos organismos vivos. Podem ter funções catalizadoras, reguladoras, transportadoras, armazenadoras, fornecedoras de apoio e proteção imunitária, além de outras mais. Podem ser descritas como polímeros lineares construídos a partir de sequências de aminoácidos, sendo essas sequências equivalentes das sequências de um mRNA, que por sua vez foi construído a partir da sequência de um gene existente no DNA.

Os aminoácidos, ou unidades monoméricas, são compostos de um grupo de amina (-NH3), um grupo carboxila (-COOH) e uma cadeia específica para cada aminoácido, chamado de radical. Existem vinte diferentes radicais comumente encontrados em proteínas, tendo variações de tamanho, forma, carga, caráter hidrofóbico e outras características (BERG; TYMOCZKO; STRYER, 2010). A FIGURA 2 mostra a estrutura química dos aminoácidos e suas respectivas classificações por grupos.

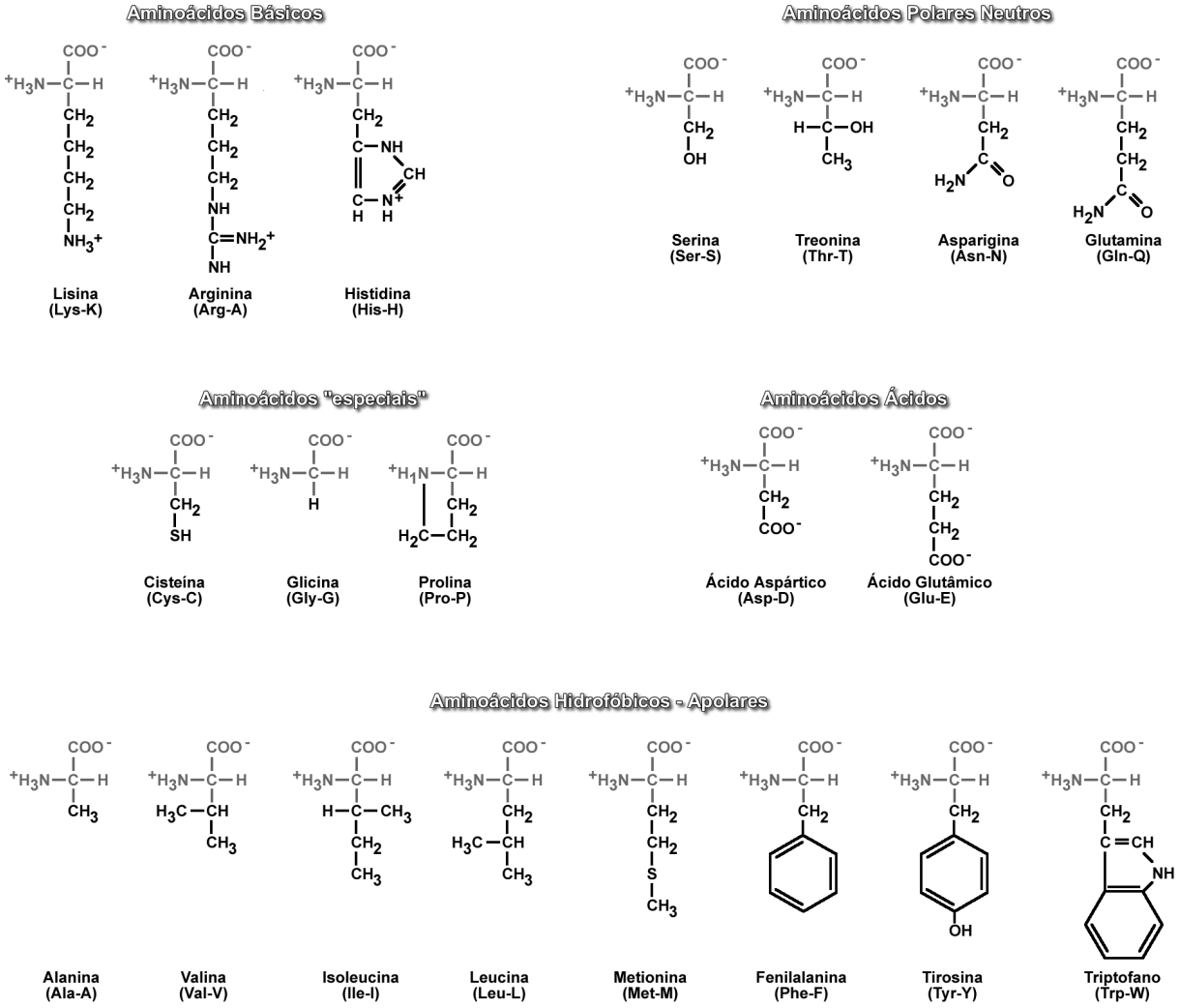

FIGURA 2 – ESTRUTURA QUÍMICA E GRUPOS DOS AMINOÁCIDOS
FONTE: O Autor.

Conhecendo as sequências de aminoácidos de uma proteína qualquer, usualmente, a homologia pode ser inferida a partir de similaridade entre essas sequências, essa similaridade podendo ser obtida através de alinhamentos entre sequências. No entanto há casos em que sequências homólogas apresentam baixa similaridade, porém suas estruturas são conservadas. Sequências de organismos diferentes podem ser comparadas, então quando dois segmentos (seja de DNA, RNA ou proteína) apresentam similaridades podemos inferir que eles partilham sequências de um ancestral comum. Quando dois segmentos possuem ancestrais comuns são chamados homólogos, e ocorrendo devido a um evento de especiação são chamados ortologos, ou caso ocorra devido a um evento de duplicação são chamados parálogos (KOONIN; GALPERIN, 2003).

2.1.3 Dogma Central da Biologia Molecular

O conceito que descreve o modo com que informações são transferidas através de um sistema biológico é chamado de Dogma Central da Biologia Molecular, e nele classificam-se as transferências gerais que representam o fluxo normal da informação biológica. Foi postulado por Francis Crick (CRICK, 1970), explicando como ocorrem esses fluxos de informações e que tais fluxos não seguem a direção inversa. Isso significa que cópias de informações são possíveis de DNA para DNA (replicação), assim como um DNA pode ter suas informações reescritas na forma de um mRNA (transcrição) e as proteínas podem ser sintetizadas a partir das informações do mRNA (tradução). Contudo é importante ressaltar que, mesmo a transcrição sendo apresentada muitas vezes de modo unidirecional, algumas vezes as cadeias de RNA agem como moldes para sínteses de cadeias de sequências complementares de DNA (WATSON et al., 2006), ou seja, pode haver um fluxo específico na direção contrária (transcrição reversa) em processos biológicos. A FIGURA 3 exemplifica o esquema desses fluxos de informações.

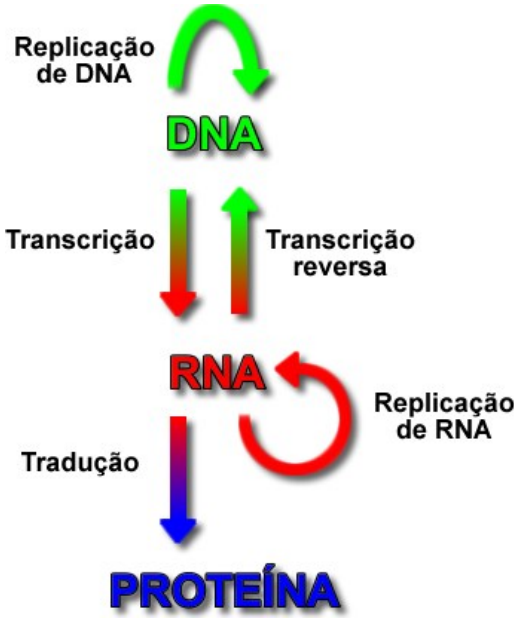

FIGURA 3 – DOGMA CENTRAL DA BIOLOGIA MOLECULAR
FONTE: O Autor.

Dessa forma, partindo da existência de vinte aminoácidos e quatro nucleotídeos do mRNA, a informação da sequência genética é determinada por tríplices de nucleotídeos ou bases, chamados de códons, que determinarão a sequência de aminoácidos que comporá a estrutura de uma proteína qualquer, agregando à proteína as várias

características que determinarão sua função no organismo. Porém nem todas as possibilidades de códons determinam o tipo de aminoácido que será ligado à sequência peptídica, existindo três combinações que resultam na parada do processo de síntese pelo ribossomo. Esses três códons específicos, conhecidos como códons de parada (*stop codons*), atuam como sinalizadores para que a sínteses seja finalizada e a proteína fique liberada do ribossomo, ou seja, num processo normal a proteína está traduzida em sua totalidade e pronta para interagir bioquimicamente no organismo. Também existe um códon específico que sinaliza a região de início da sequência da proteína, chamada de códon de início (*start codon*), que além de determinar a inclusão de um aminoácido específico na síntese de proteína (NELSON; COX, 2006). A TABELA 1 indica cada sequência dos códons de mRNA e o nome aminoácido resultante, bem como suas duas formas de abreviação.

TABELA 1 – CODIFICAÇÃO DE AMINOÁCIDOS

PRIMEIRA BASE	SEGUNDA BASE U	SEGUNDA BASE C	SEGUNDA BASE A	SEGUNDA BASE G	TERCEIRA BASE
U	UUU Fenilalanina (Phe – F)	UCU Serina (Ser – S)	UAU Tirosina (Tyr – Y)	UGU Cisteína (Cys – C)	U
	UUC Fenilalanina (Phe – F)	UCC Serina (Ser – S)	UAC Tirosina (Tyr – Y)	UGC Cisteína (Cys – C)	C
	UUA Leucina (Leu – L)	UCA Serina (Ser – S)	UAA ***Stop codon***	UGA ***Stop codon***	A
	UUG Leucina (Leu – L)	UCG Serina (Ser – S)	UAG ***Stop codon***	UGG Triptofano (Trp – W)	G
C	CUU Leucina (Leu – L)	CCU Prolina (Pro – P)	CAU Histidina (His – H)	CGU Arginina (Arg – R)	U
	CUC Leucina (Leu – L)	CCC Prolina (Pro – P)	CAC Histidina (His – H)	CGC Arginina (Arg – R)	C
	CUA Leucina (Leu – L)	CCA Prolina (Pro – P)	CAA Glutamina (Gln – Q)	CGA Arginina (Arg – R)	A
	CUG Leucina (Leu – L)	CCG Prolina (Pro – P)	CAG Glutamina (Gln – Q)	CGG Arginina (Arg – R)	G
A	AUU Isoleucina (Ile – I)	ACU Treonina (Thr – T)	AAU Asparagina (Asn – N)	AGU Serina (Ser – S)	U
	AUC Isoleucina (Ile – I)	ACC Treonina (Thr – T)	AAC Asparagina (Asn – N)	AGC Serina (Ser – S)	C
	AUA Isoleucina (Ile – I)	ACA Treonina (Thr – T)	AAA Lisina (Lys – K)	AGA Arginina (Arg – R)	A
	AUG Metionina (Met – M) ***Start codon***	ACG Treonina (Thr – T)	AAG Lisina (Lys – K)	AGG Arginina (Arg – R)	G
G	GUU Valina (Val – V)	GCU Alanina (Ala – A)	GAU Ácido aspártico (Asp – D)	GGU Glicina (Gly – G)	U
	GUC Valina (Val – V)	GCC Alanina (Ala – A)	GAC Ácido aspártico (Asp – D)	GGC Glicina (Gly – G)	C
	GUA Valina (Val – V)	GCA Alanina (Ala – A)	GAA Ácido glutâmico (Glu – E)	GGA Glicina (Gly – G)	A
	GUG Valina (Val – V)	GCG Alanina (Ala – A)	GAG Ácido glutâmico (Glu – E)	GGG Glicina (Gly – G)	G

Três bases correspondem ao aminoácido a ser adicionado na cadeia peptídica.
FONTE: O Autor.

Assim, cada códon, com exceção dos *stop codons* (UAA, UAG e UGA), determinam um aminoácido a ser acrescido na sequência da montagem de uma proteína,

e o fato de alguns códons resultarem na adição do mesmo tipo de aminoácido faz o código genético ser chamado degenerado.

2.1.4 Processo de tradução

O gene é, por definição bioquímica, toda região de DNA que codifica a sequência primária de um produto gênico final, podendo ser um polipeptídeo (proteína) ou um RNA. Para determinar como um gene é expresso é necessário analisar como funcionam os fluxos das informações determinadas pelo dogma central da biologia molecular (NELSON; COX, 2006).

No DNA também estão contidos outros segmentos que possuem papéis reguladores. Essas sequências reguladoras agem como sinais que podem denotar o início ou fim das regiões gênicas, além de influenciar na transcrição ou funcionar como pontos de início para a replicação ou recombinação. Isso evidencia como os genes podem ser expressos de maneiras diferentes, gerando múltiplos produtos a partir de um mesmo segmento de DNA (NELSON; COX, 2006).

O DNA, quando sofre o processo de transcrição, tem sua dupla hélice envolta por uma molécula compostas de várias subunidades, conhecida como RNA-polimerase, que desenrola trechos de suas sequências nucleotídicas. A RNA-polimerase utiliza uma das duas fitas do DNA como molde para a síntese complementar progressiva de um RNA, através do pareamento das bases, produzindo o mRNA. Esse mRNA é produzido de modo diferente entre células eucarióticas e procarióticas, sendo o processo mais complexo ocorrente nos organismos eucariotos (WATSON et al., 2006), foco deste trabalho.

Nas células procariontes o mRNA sintetizado já está pronto para ser utilizado no processo de tradução. Já em células eucariontes o RNA precisa sofrer uma série de outros eventos para sua maturação, ou seja, para estar pronto para atuar como um mRNA. Entre esses eventos necessários pode-se destacar o *splicing* (clivagem) de trechos do RNA não maduro, gerando fragmentos que são classificados como íntrons ou éxons. Os íntrons são as sequências do RNA que não são responsáveis diretamente pelo processo de tradução, agindo de outras formas nos demais eventos biológicos. Já os éxons são os trechos do mesmo RNA que serão unidos entre si por outras moléculas formando uma nova sequência contínua de RNA, resultando no mRNA pronto para ser traduzido para a forma de um polipeptídeo (WATSON et al., 2006). A FIGURA 4

exemplifica basicamente os processos de transcrição e tradução nos organismos eucarióticos.

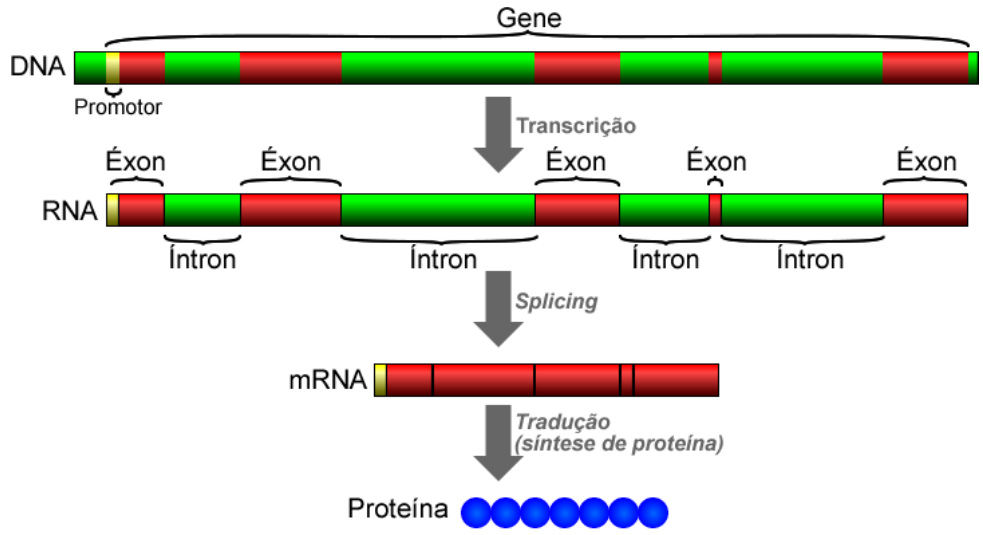

FIGURA 4 – FLUXO DE PROCESSOS TRANSCRICIONAIS E TRADUCIONAIS
FONTE: O Autor.

O processo de tradução de um mRNA exige quatro principais agentes: o próprio mRNA; moléculas adaptadoras (tRNA); as enzimas que ligam os aminoácidos nos tRNA; e o ribossomo (basicamente o rRNA) que une sequencialmente os aminoácidos que formarão uma proteína completa. As regiões que codificam uma proteína, são compostas por uma sucessão contínuas de códons chamada de fase aberta de leitura ou ORF (*open-reading frame*), os quais definem uma única proteína, tendo início e fim localizados no mRNA (WATSON et al., 2006).

Um mRNA contém pelo menos uma ORF e sua quantidade pode variar para eucariotos e procariotos. Nas células eucarióticas o mRNA possui uma única ORF e são chamados de mRNA monocistrônicos, e qualquer mRNA que possua mais de uma ORF é chamado de mRNA policistrônico (WATSON et al., 2006). A FIGURA 5 ilustra o processo de tradução, leitura das informações do mRNA pelo ribossomo e montagem da sequência de aminoácidos.

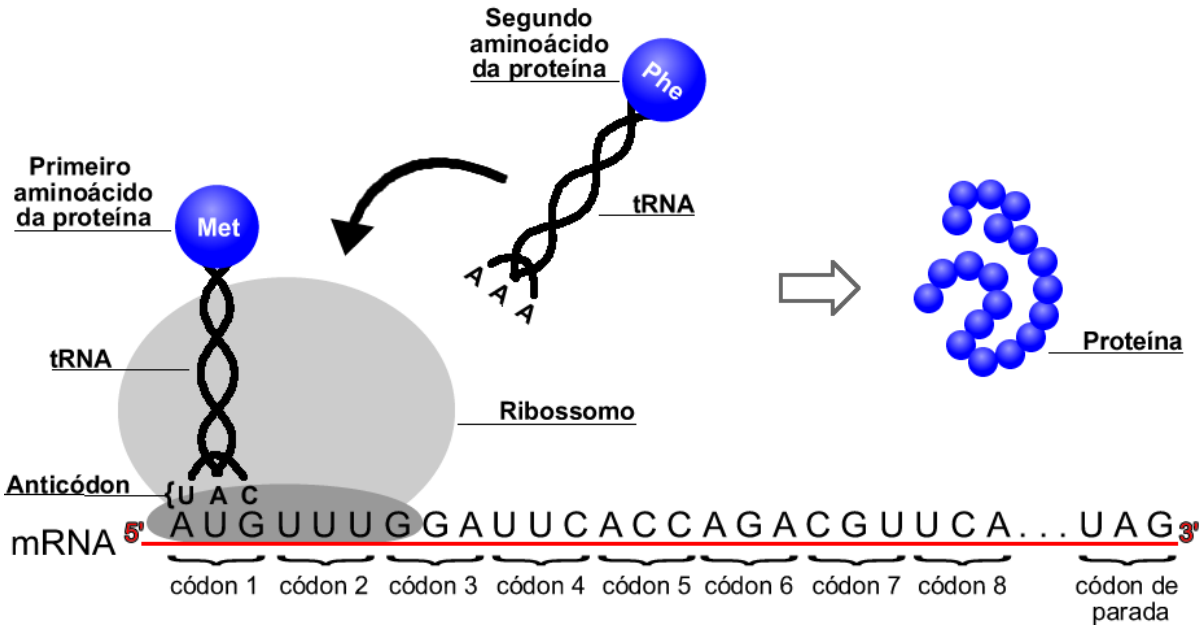

FIGURA 5 – ILUSTRAÇÃO DO PROCESSO DE TRADUÇÃO

FONTE: O Autor.

Conforme a ilustração acima, inicia-se a tradução na extremidade 5' do mRNA, prosseguindo códon por códon até a extremidade 3'. As células eucarióticas usam o códon “AUG” como códon de inciação. Ele também tem a função de definir o primeiro aminoácido a incorporar à sequência polipeptídica, assim como os códons de parada (UAG, UGA e UAA) sinalizam o final da ORF, terminando a síntese do polipeptídeo mas sem adicionar um aminoácido “final” (WATSON et al., 2006).

2.1.5 Sequenciamento

Para obtermos as sequências de nucleotídeos (DNA e RNA) ou de aminoácidos (proteínas), como informação, é necessário realizar processos bioquímicos de sequenciamento em laboratório, para o qual existem diferentes métodos que proporcionam tamanhos, qualidades e coberturas diferentes das sequências objetos de estudo.

O primeiro sequenciamento automático foi desenvolvido através da metodologia de Sanger (SANGER; NICKLEN, 1977), que tinha elevado custo e por isso inviabilizava projetos menores. Mas novos métodos foram desenvolvidos a partir de 2004, produzindo novos sequenciadores para o mercado equanto reduziam-se o custo e tempo. Essas

tecnologias ficaram conhecidas como sequenciamento de nova geração (*Next generation sequencing* – NGS) e alavancaram muitas outras pesquisas de genética e genômica. Os métodos mais avançados e atualmente utilizados proporcionam uma gigantesca diferença, positiva para o pesquisador, em relação ao custo, qualidade e tempo dos dados obtidos num processo de sequenciamento (QUAIL et al., 2012).

O primeiro sequenciamento não viral realizado foi da bactéria *Haemophilus influenzae* (FLEISCHMANN et al., 1995), seguidos pelos sequenciamentos de bactérias como a *Escherichia coli* (BLATTNER et al., 1997) e a *Mycobacterium tuberculosis* (COLE et al., 1998). A partir de então, junto com o sequenciamento do genoma humano (LANDER et al., 2001), houve um grande aumento na quantidade de genomas sequenciados e registrados em banco de dados, ultrapassando-se a marca de mil genomas completamente sequenciados em 2010. Desde então a quantidade de genomas depositados em bancos de dados cresce radicalmente a cada ano, principalmente pelos desenvolvimentos e custos decrescentes das tecnologias desenvolvidas. Na FIGURA 6 é mostrado o gráfico dos custos de sequenciamento dos genomas com o passar dos anos.

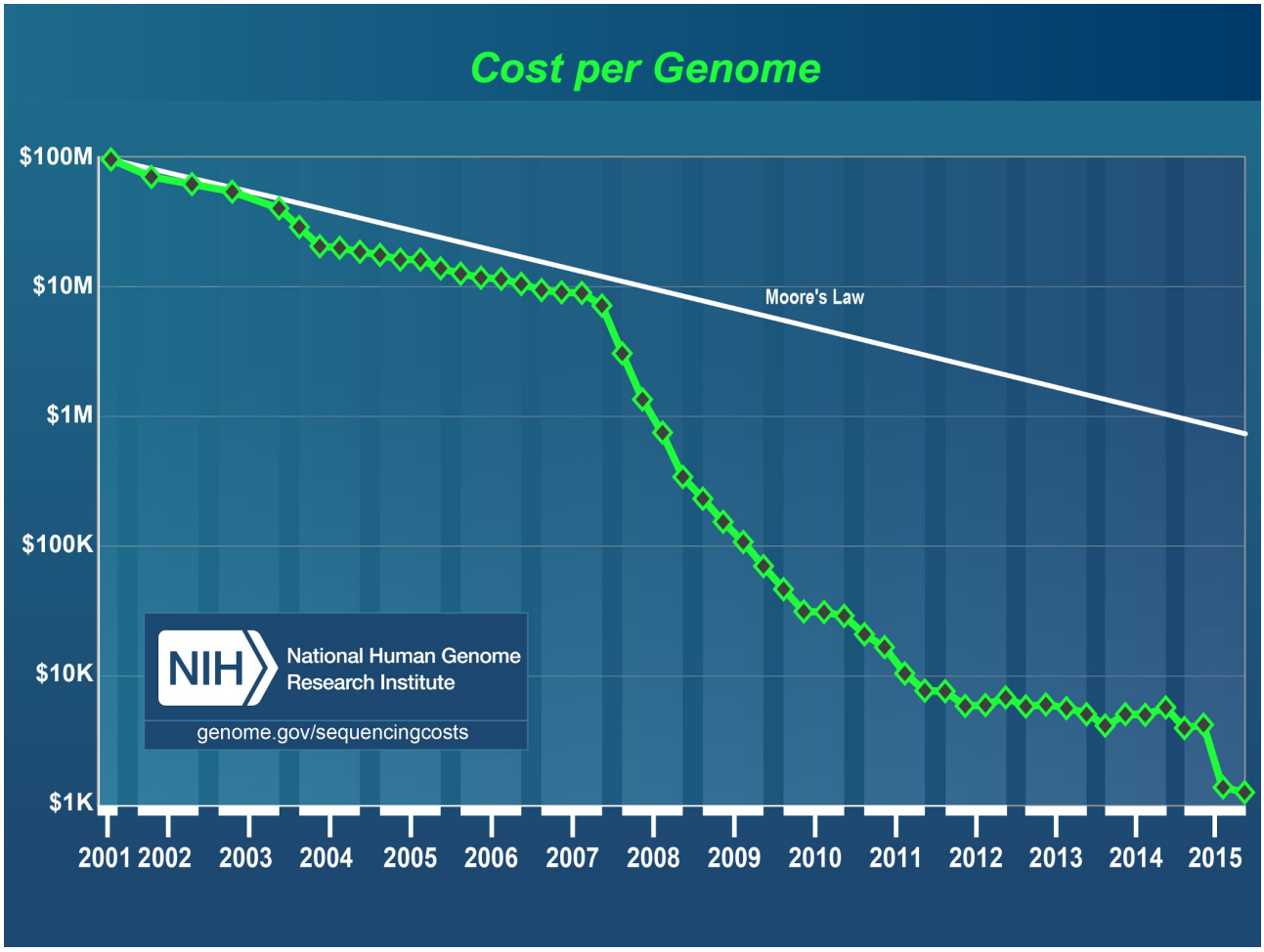

FIGURA 6 – CUSTO DE SEQUENCIAMENTO DE GENOMAS

Os valores em dólar mostram a queda no custo de sequenciamentos com o passar dos anos.

FONTE: NIH (http://genome.gov, 2016).

O gráfico acima esclarece a forte queda dos custos de sequenciamento dos genomas, acentuada a partir de 2007 . Observa-se também que em 2015 houve uma nova queda nesses custos, o que viabiliza atuais pesquisas e pequenos projetos que necessitem de sequenciamentos.

2.1.6 Montagem de genomas

A montagem de uma sequência refere-se ao processo de unir grandes quantidades de pequenos trechos obtidos como dados de computador pelo processo de sequenciamento. A análise e união dessas sequências curtas, também chamadas de leituras ou *reads*, tem como objetivo obter a sequência original completa. Essas

montagens utilizam algoritmos que geralmente trabalham tentando unir as leituras entre si, encaixando-as de modo a se obter as sequências consenso, também chamadas de *contigs,* resultando nos *scaffolds* (VIALLE, 2013), como mostrados na FIGURA 7.

FIGURA 7 – EXEMPLIFICAÇÃO DOS *READS*, *CONTIGS* E *SCAFFOLDS*.
FONTE: Ricardo Vialle (VIALLE, 2013).

A montagem de genomas gera um problema computacional complexo, principalmente nos sequenciamentos de genomas que possuem grandes quantidades de sequências repetidas, podendo ser repetições com centenas de nucleotídeos e em diferentes regiões do genoma. A união dos contigs é chamado de *scaffolds* e os resultados obtidos com os *contigs* e *scaffolds* é chamado de esboço (do inglês draft) do genoma. A etapa de finalização (do inglês *finishing*) tenta corrigir a saída fragmentada do montador, erros nos contigs, além do fechamento de ausências, chamados de *gaps*, validando a montagem (VIALLE, 2013).

2.1.7 Anotação de sequências

Obter informações estruturais e funcionais sobre uma ou várias sequências de um

genoma é o objetivo do processo de anotação (STEIN, 2001). É preciso integrar análises computacionais aos dados biológicos auxiliares, além de algumas outras integrações para obter-se a maior quantidade possível de dados úteis nas inferências dessas sequências (LEWIS et al., 2000).

Nesse processo de anotação genômica, um DNA sequenciado é analisado e documentado por meio da identificação dos vários sítios e segmentos envolvidos no funcionamento do genoma (ROUZÉ et al., 1999). Consiste em duas etapas principais, sendo a predição dos genes e a agregação de informações verificadas sobre eles. A etapada da predição, também conhecida como estrutual, proporciona a identificação de elementos como ORFs e localização de motivos reguladores. A segunda, também chamada de funcional, identifica, por exemplo, as funções bioquímicas, interações e expressões gênicas. Essas duas etapas podem envolver experimentos biológicos e análises in silico (VIALLE, 2013).

Com a chegada do *Next Generation Sequencing* (NGS), é possível produzir *drafts* em pouco tempo, ocasionando a necessidade da anotação de genomas de forma rápida. As ferramentas de anotação automática realizam todas as etapas por meio da análise computacional, enquanto a anotação manual necessita da curadoria realizada por um especialista humano. Dessa forma as anotações automatizadas oferecem ganhos em tempo e custo comparados ao processo manual (PETTY, 2010). Porém, ferramentas de anotação automáticas geralmente apresentam muitas diferenças, sendo preciso uma avaliação manual final sobre os resultados obtidos (BAKKE et al., 2009). Então, na prática, devem trabalhar juntas ambas as abordagens, complementando-se a fim de obter-se resultados com mais qualidade. Na FIGURA 8 é ilustrada as etapas de anotação de genes.

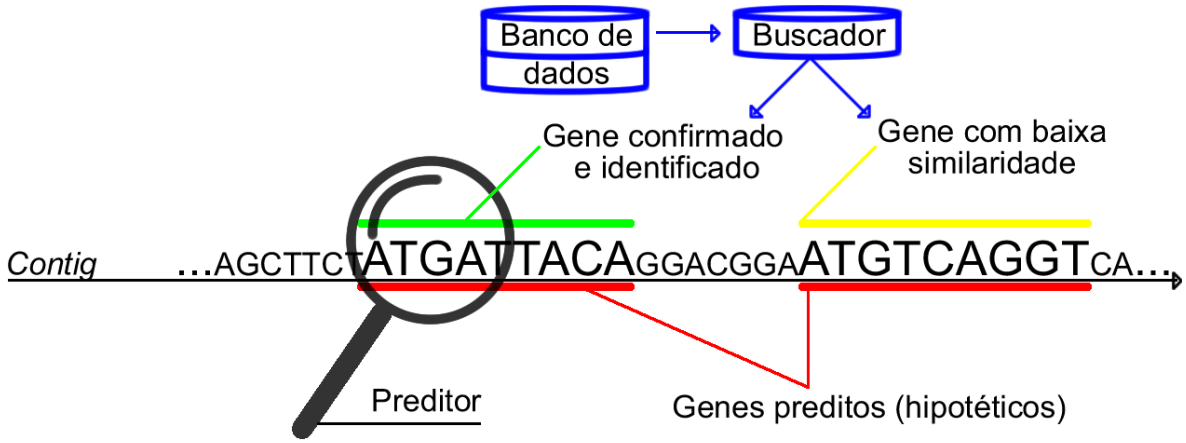

FIGURA 8 – ILUSTRAÇÃO DA ANOTAÇÃO DE GENES

FONTE: O Autor..

Algoritmos de buscadores como o BLAST e outras ferramentas de comparação participam do nível básico de anotação, buscando por similaridades em bancos de dados com informações anotadas em outros genomas. Outras informações podem ser agregadas na anotação como função e organismo proveniente, e quanto maior a quantidade de sequências e informações melhor será a tarefa do curador para resolver discrepâncias. Assim, alguns bancos de dados baseiam-se em contextos de genomas, mensuração de similaridades, dados experimentais e integração com outras fontes de dados com o objetivo de aumentar o volume de dados para as buscas.

2.2 ARMAZENAMENTO E OBTENÇÃO DE INFORMAÇÕES BIOLÓGICAS

Existem basicamente dois tipos de bancos de dados biológicos disponíveis para utilização. Nos bancos primários são depositadas informações experimentais com pouca ou sem interpretação, onde não há uma análise cuidadosa com relação as sequências. Este é o caso dos bancos GenBank e EMBL (STOESSER et al., 2002), por exemplo. Os bancos de dados secundários já possuem uma compilação e interpretação dos dados de entrada de forma que podem ser obtidas informações mais representativas e atualizadas. Esses são os bancos de dados curados, como por exemplo o SwissProt e o TrEMBL (BAIROCH e APWEILER, 1997).

Nos bancos de dados biológicos são depositados muitas informações obtidas em pesquisas sobre estrutura e função de moléculas de diversos organismos. Existem dezenas de bancos disponíveis, sendo que entre eles existem diferenças de tipo, formato e modo de acessso aos registros. Destacam-se alguns como Genbank, Uniprot, EMBL, INSDC, EBI, DDBJ entre outros. Neste trabalho serão abordados dois bancos de proteínas, cujo foram utilizados na realização dos testes e avaliação do Sila Eukariotic, sendo eles o SwissProt e o GenBank non-redundant (NR).

2.2.1 GenBank

É um banco de dados público, criado e compartilhado pelo National Center for Biotechnology Information (NCBI), uma divisão da National Library Medicine (NLM), situado no campus do US National Institutes of Health (NIH) em Bethesda (BENSON et

al., 2009). Com a colaboração do European Bioinformatics Institute (EBI) do European Molecular Biology Laboratory (EMBL) e do DNA Data Bank of Japan (DDBJ), que constituem o International Nucleotide Sequence Database Collaboration (INSDC), as informações são adicionadas ao GenBank. Esses dados são disponibilizados gratuitamente pelo NCBI em arquivos de texto com formatos padronizados, oferecendo também ferramentas para análises dessas informações. Cada registro possui descrição taxonomia, regiões codificantes, referências bibliográficas, tradução das proteínas e outras. Esse padrão de estrutura é definido junto ao EMBL, recebendo o nome de GenBank. Exemplificamos o formato GenBank com a FIGURA 9.

```
LOCUS       LISOD                    756 bp    DNA     linear   BCT 30-JUN-1993
DEFINITION  Listeria ivanovii sod gene for superoxide dismutase.
ACCESSION   X64011 S78972
VERSION     X64011.1  GI:44010
KEYWORDS    sod gene; superoxide dismutase.
SOURCE      Listeria ivanovii
  ORGANISM  Listeria ivanovii
            Bacteria; Firmicutes; Bacillales; Listeriaceae; Listeria.
REFERENCE   1  (bases 1 to 756)
  AUTHORS   Haas,A. and Goebel,W.
  TITLE     Cloning of a superoxide dismutase gene from Listeria ivanovii by
            functional complementation in Escherichia coli and characterization
            of the gene product
  JOURNAL   Mol. Gen. Genet. 231 (2), 313-322 (1992)
  MEDLINE   92140371
REFERENCE   2  (bases 1 to 756)
  AUTHORS   Kreft,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-APR-1992) J. Kreft, Institut f. Mikrobiologie,
            Universitaet Wuerzburg, Biozentrum Am Hubland, 8700 Wuerzburg, FRG
FEATURES             Location/Qualifiers
     source          1..756
                     /organism="Listeria ivanovii"
                     /strain="ATCC 19119"
                     /db_xref="taxon:1638"
                     /mol_type="genomic DNA"
     RBS             95..100
                     /gene="sod"
     gene            95..746
                     /gene="sod"
     CDS             109..717
                     /gene="sod"
                     /EC_number="1.15.1.1"
                     /codon_start=1
                     /transl_table=11
                     /product="superoxide dismutase"
                     /db_xref="GI:44011"
                     /db_xref="GOA:P28763"
                     /db_xref="InterPro:IPR001189"
                     /db_xref="UniProtKB/Swiss-Prot:P28763"
                     /protein_id="CAA45406.1"
                     /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVS
                     GHAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLK
                     AAIESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPV
                     LGLDVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
     terminator      723..746
                     /gene="sod"
ORIGIN
        1 cgttatttaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat
       61 gtaatttctt ..........
//
```

FIGURA 9 – EXEMPLO DO FORMATO GENBANK

FONTE: NCBI (2015).

O GenBank possui uma versão de banco de dados, utilizada neste projeto,

chamada de Não Redundante (NR), do inglês *Non-Redundant*, em que os registros nele contidos não se repetem e são a somatória de todas as proteínas já estudadas (curadas) com todas as possivelmente existentes (hipotéticas). No ano de 2014 o banco NR contava com mais de 54 milhões de sequências, e no final de 2015 essa somatória já passava dos 78 milhões. Esse banco também pode ser obtido em formato muilt-Fasta, o qual será explicado no tópico 2.2.3.

2.2.2 SwissProt

O catálogo *Universal Protein Resource* (UniProt) é um repositório central de sequências de proteínas. Dele ramifica-se um banco de dados não redundante e com referências cruzadas de outros 24 bancos, contendo somente sequências de proteínas curadas, chamado de SwissProt. Foi estabelecido em 1986 sendo mantido em colaboração pelo Department of Medical Biochemistry of the University of Genova e EMBL Data Library (BAIROCH e APWEILER, 1997).

Como o SwissProt contém somente sequências curadas, manualmente anotadas, trata-se de um banco de dados relativamente pequeno, se comparado ao GenBank, possuindo no ano de 2016 pouco mais de 500 mil sequências depositadas. Também é um banco de dados biológicos público, sendo disponibilizado gratuitamente pelo UniProt para utilização. Esse banco também pode ser obtido em formato multi-Fasta.

2.2.3 Fasta

Tanto o banco GenBank-NR quanto o SwissProt, assim como sequências genômicas, podem ser obtidos no formato multi-Fasta, ou seja, duas ou mais sequências no formato Fasta contido num mesmo arquivo de computador. Este é o formato mais comum para representar sequências biológicas, sejam nucleotídicas ou de aminoácidos. O Fasta teve origem junto ao software de mesmo nome (LIPMAN; PEARSON, 1985), tornando-se um padrão na Bioinformática. Neste formato uma sequência possui um cabeçalho de informações iniciado pelo caractere “>” (maior que), seguido por identificações alfanuméricas para nome, banco de dados de origem, organismo e outras. A próxima linha deve conter necessariamente a sequência biológica correspondente, ou

seja, condizente ao cabeçalho de identificação. Um arquivo fasta pode conter várias sequências seguindo o mesmo padrão, nesse caso, quando temos mais de uma sequência no mesmo arquivo Fasta, chamamos esse arquivo/formato de multi-Fasta, sendo que cada sequência deve ter obrigatoriamente uma identificação única.

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

FIGURA 10 – EXEMPLO DO FORMATO FASTA
FONTE: NCBI (2015).

2.3 ALGORITMOS E SOFTWARES RELACIONADOS

2.3.1 Preditores de genes

A predição automática de genes é feita por softwares que buscam identificar regiões codificantes (que geram RNA) através de evidências obtidas por técnicas como programação dinâmica, inteligência artificial (algoritmos genéticos, redes neurais artificiais) e métodos estatísticos. A TABELA 2 elenca alguns dos softwares preditores de genes eucarióticos, e na sequência uma breve descrição de alguns deles.

TABELA 2 – LISTA DE SOFTWARES PREDITORES DE GENES EUCARIÓTICOS

Software	Ano	Testes originais de sequências	Plataforma	Linguagem	Técnica global	Fonte
AraNet	2011	*Arabidopsis thaliana*	Web	Java	NN	HWANG et al., 2011.
BioPixie	2005	*Saccharomyces cerevisiae*	Web, Linux, Windows	PHP, Perl	EP	MYERS et al., 2005.
EuGène	2014	*Arabidopsis* e plantas	Web, Linux	C++	HMM	SALLET et al., 2014.
Fgeneh	1994	Humanos, mosca da fruta, Drosophila e plantas	Web	C++	HMM	SALAMOV et al., 2000.
FuncAssociate	2003	Humanos, Drosophila, Arabidopsis e ratos	Web, Linux, Windows	C, Perl	GOA	BERRIS et al., 2003.
Gene Id	1991	Vertebrados e plantas	Unix	C	EP	GUIGÓ et al., 1992.
GeneMANIA	2010	Humanos, ratos e vermes	Web	Java	RME	FARLEY et al., 2010.
GeneMark ES	2014	Humanos, ratos, Drosophila e Gallus	Linux, Unix, Solaris	C	NN e HMM	LOMSADZE et al., 2014.
GeneMarkHMM	1997	Archeas e bactérias	Linux, Mac, Solaris	C	HMM	LUKASHIN et al., 1998.
GeneParser	1995	Vertebrados e plantas	Windows, Linux, Mac	Perl	NN	SNYDER et al., 1995.
Genezilla	2004	Plantas e fungos	Windows, Linux, Mac	C/C++	HMM e GHMM	MAJOROS et al., 2004.
Genie	1996	Humanos	*	*	DP, NN e GHMM	KULP et al., 1996.
GENSCAN	1997	Humanos	Web	*	HMM e MDD	BURGE et al., 1997.
GRAIL II	1994	Humanos e ratos	Web e Linux	*	NN e DP	XU et al., 1994.
GRPL	2000	Humanos, Drosophila e Arabidopsis	Sun Solaris	*	DP e GHMM	HOOPER, et al., 2000.
HMMgene	1997	Humanoss, Drosophila e dicotiledônias	Web	*	HMM	KROGH, 2000.
mGene	2009	Eucariotos*	Web, Linux	C++	NN e HSM-SVM	SCHWEIKERT, et al., 2009.
MORGAN	1998	Vertebrados, *Drosophila* e dicotiledônias	**	*	DP e MM	SALZBERG, et al., 1998.
MZEF	1997	Humanos, ratos e Arabidopsis	Linux	*	QDA	ZHANG, 1997.
Phat	2001	Humanos e Plasmodium virax	Linux	Perl	GHMM	WIRTH, 2001.
TigrScan	2004	Plantas e fungos	Linux e Mac	C/C++	HMM e GHMM	MAJOROS et al., 2004.
TWAIN	2004	Fungos	Linux, Unix	C/C++	HMM e GPHMM	MAJOROS et al., 2005.
TwinScan	2004	Vegetais	Windows, Linux, Mac	C++	DP e IMM	ALLEN et al., 2004.
VEIL	1997	Vertebrados e *Drosophila*	**	*	HMM	HENDERSON, et al., 1997.
VIRGO	2006	Humanos, fungos e Drosophila	Web **	Java	FLN	MASSOUNI, et al., 2006.

* Não informado no artigo original. ** Projeto descontinuado ou servidor não encontrado.
Técnica Global: *GeneOntology Atributes* (GOA), Estatística probabilística (EP), Redes e Múltiplas Evidências (RME), *Neural Network* (NN), *Dinamic Programming* (DP), *Maximal Dependence Decomposition* (MDD), *Markov Model* (MM), *Interpoled Markov Model* (IMM), *Hidden Markov Model* (HMM), *Generalized Hidden Markov Model* (GHMM), *Generalized pair Hidden Markov Model* (GPHMM), *Hidden semi-Markov Suport Vector Machines* (HSM-SVM), *Quadratic Discriminant Analysis* (QDA), *Functional Linkage Network* (FLN).
FONTE: O Autor.

É possível observar, na TABELA 2, que os preditores utilizam vários tipos de abordagens para realizar seu trabalho (técnica global), e alguns deles inclusive combinam mais de uma técnica no objetivo de melhorar a predição de genes. Apesar dos preditores serem desenvolvidos geralmente para um alvo ou grupo de organismos, nada impede deles encontrarem genes em genomas de organismos distintos, contudo é possível que sua eficácia e sensibilidade sejam comprometidas quando utilizados para genomas fora de seus escopos. Explicaremos na sequência o funcionamento básico de alguns dos softwares preditores.

2.3.1.1 GeneMark-ES

O GeneMark é um grupo de programas para predições gênicas desenvolvido em

Atlanta (EUA) no *Georgia Institute of Technology*. A sua versão “ES” (*Eukaryotic Self-training*) é disponibilizada para predição de genes em genomas eucarióticos. Ele realiza o auto treinamento não supervisionado em base do genoma de entrada a fim de obter melhor resultado na localização dos genes e seus éxons. Também pode ser utilizado o modo de busca específicas para genes de fungos. No final de sua execução é gerado um arquivo no formato GTF (semelhante ao GFF), contendo as coordenadas e outras informações dos possíveis genes e éxons encontrados (LOMSADZE et al., 2005). É um software gratuito para fins não comerciais, que depende de uma chave de ativação que deve ser solicitada previamente em seu site oficial (http://exon.gatech.edu/GeneMark/) para seu funcionamento, válida por doze meses consecutivos . Atualmente seu algoritmo está disponível para execução nos sistemas operacionais Linux e MacOS.

2.3.1.2 TigrScan

Preditor que tem seu algoritmos escrito nas linguagens C e C++, disponível em código aberto. Ele baseia-se no Modelo Oculto de Markov para a predição de genes eucarióticos, podendo ser retreinado, ao contrário de alguns outros preditores. Também pode ser reconfigurado e inclui vários submodelos probabilísticos que podem ser combinados de forma idependente, tais como árvores de dependência e Modelo Interpolado de Markov (MAJOROS et al., 2004).

2.3.1.3 Jigsaw

É um sistema preditor automático de genes que pode utilizar multiplas fontes de evidências contidas em arquivos, sendo seus resultados geralmente muito próximo aos da curagem humana. Ele processa os pesos relativos de diferentes linhas de evidências, usando estatísticas geradas a partir de um conjunto de treinamentos e então combina as evidências com técnicas de programação dinâmica (ALLEN; SALZBERG, 2005). Seu algoritmo de código aberto, escrito na linguagem C++, necessita ser compilado para ser executado.

2.3.1.4 AraNet

Trata-se de um software de predição de função de genes voltado para genomas de origem vegetal, em especial a planta *Arabidopsis*. Sua proposta visa melhorar a predição identificando genes associados à características de plantas. É altamente preditivo para diversas vias biológicas. Além disso, o AraNet fornece uma ferramenta para web, com a qual os usuários podem descobrir eficientemente novas funções de genes utilizando redes de funções (HWANG et al., 2011).

2.3.2 Comparadores de Sequências

Os softwares comparadores de sequências, também conhecidos como buscadores de similaridade, fazem pesquisas em bancos de dados biológicos a fim de obter a(s) sequência(s) mais parecidas num banco de dados para as sequências candidatas (*query*). Os buscadores podem utilizar técnicas de programação dinâmica, alinhamento, indexação e métodos estatísticos. Serão apresentados nessa seção alguns desses programas.

2.3.2.1 BLAST

A ferramenta BLAST (*Basic Local Alignment Search Tool*) realiza buscas e comparações de sequências tanto de proteínas quanto de DNA. É um dos programas mais utilizados na Bioinformática, apresentando alta sensibilidade e desempenho superior se comparado com algoritmos que utilizam programação dinâmica. Para suas comparações o BLAST realiza alinhamento local, ou seja, busca por similaridades em regiões específicas de uma sequência. A busca de sequências similares é feita a partir de uma sequência de entrada (*query*) a qual é comparada com todas as sequências contidas num banco de dados. Seus resultados são confiáveis e podem ser utilizados em diferentes cenários (ALTSCHUL et al., 1990).

O BLAST aborda um sistema de sementes para realizar a pré-seleção dos locais com possível alinhamento, após, os alinhamentos são estendidos seguindo um sistema de escore (pontuação) de similaridade. Esse sistema de escore pode ser reconfigurado para aumentar a sensibilidade das comparações, o que poderá aumentar seu tempo de resposta. Basicamente seu algoritmo cria uma lista de sequências (palavras) que serão

comparadas com o banco de dados, obtendo-se uma lista de sequências com regiões similares, então o alinhamento é extendido para as sequências encontradas, expandindo até que sua pontuação atinja um limite mínimo de similaridade.

Essa ferramenta pode ser utilizada em ambiente web e também executada em ambiente local, além do pesquisador poder escolher o banco de sequências de sua preferência.

2.3.2.2 BLAT

Do acrônimo "*BLAST-Like Alignment Tool*", é uma ferramenta alternativa semelhante ao BLAST e utiliza técnicas de alinhamento e indexação. Realiza comparações tanto de sequências nucleotídicas como de aminoácidos, apresentando ótima acurácia e chegando a ser 500 vezes mais rápido que as ferramentas populares de alinhamento (KENT, 2002). A velocidade do BLAT decorre de um índice de todos os K-mers (sementes) não sobrepostos, essa indexação, carregado na memória RAM do computador, precisa ser calculada apenas uma vez para cada montagem de sequência(s).

O BLAT utiliza o índice para encontrar regiões que sejam homólogas à(s) sequência(s) de consulta, realizando um alinhamento entre essas regiões homólogas. Ele une essas regiões alinhadas, muitas vezes éxons, em alinhamentos maiores, tipicamente genes. Finalmente são revisados pequenos éxons internos possivelmente perdidos na primeira fase e ajustados os gaps que possuem locais de splicing canônico sempre que possível (KENT, 2002). Pode-se escolher o banco de dados a ser consultado e está disponível para utilização nos sistemas operacionais Linux e MacOS.

2.3.2.3 RAFTS3

Ferramenta rápida para buscas de similiraridades entre sequências (*Rapid Alignment-Free Tool for Sequences Similarity Search*), o RAFTS3 não utiliza técnicas de alinhamentos (livre de alinhamento). Suas abordagens gerais para a busca e comparação são a indexação de dados e métodos estatísticos.

O primeiro passo, para o RAFTS3 realizar as buscas por similaridade, é a

indexação do banco de dados num formato específico, chamado de DBSTRUCT. A indexação consiste basicamente em criar um arquivo com índices para todas as sequências do banco e gerar uma matriz binária de co-ocorrência de aminoácidos (BCOM), representando a assinatura de cada sequência. Uma vez que o banco é indexado, o mesmo está pronto para participar dos processos de pesquisa, podendo ser utilizado quantas vezes for necessário, ou seja, não necessita de reindexação, exceto se o pesquisador desejar atualizá-lo. O DBSTRUCT permite o acesso direto do RAFTS3 ao banco de dados, de modo que a consulta da similaridade é feita através do arquivo de índices e posteriormente recupera-se a sequência no banco de sequências (VIALLE et al., 2016).

Como seu algoritmo foi escrito na linguagem MatLab, ele necessita da instalação do pacote do compilador MCR (Matlab Compiler Runtime). Esse pacote é gratuito e está disponível para download para os sistemas operacionais Windows, Linux e MacOS de 32 e 64 bits.

2.3.3 Anotadores Automáticos para Eucariotos

Consideramos como anotadores automáticos os algoritmos que fazem o trabalho de identificar regiões em uma sequência de DNA. Resumidamente esses softwares devem incorporar dados de predição, avaliar evidências ou assumir por similaridades, entre outros processos, utilizando bancos de dados com sequências biológicas para a confirmação e/ou correção de coordenadas dos genes, resultando então numa anotação evidencialmente fundamentadas e com suas regiões devidamente identificadas e/ou associadas (conceito do autor). Serão explicados brevemente alguns anotadores disponíveis para utilização.

2.3.3.1 ENCODE

O ENCODE é um algoritmo que utiliza um conjunto de softwares para fazer a anotação automática baseada em alinhamentos. Os softwares Fgenesh e o Fgenesh+, baseados no Modelo Oculto de Markov (HMM), são os preditores responsáveis pela marcação dos possíveis genes. Seus resultados contam também com os algoritmos

“Est_map” e “Prot_map” que, respectivamente, fazem mapeamentos de ESTs e proteínas na sequência genômica, possibilitando o reconhecimento também de promotores. As buscas por similaridades, a fim de verificar e confirmar a identidade dos genes, são executadas pelo algoritmo BLAST (SOLOVYEV et al., 2006).

2.3.3.2 DOGMA

O DOGMA (Dual Organellar Genome Annotator) é uma ferramenta para anotação automática de genomas de organelas (cloroplastos e mitocôndrias). Seu procedimento conta com a utilização do algoritmo BLAST para buscas de similaridades em bancos de dados. A ferramenta também provê uma interface gráfica baseada em WEB para a visualização, edição das anotações e o envio dessas anotações através do Sequin, software para submissão de sequências anotadas para o GenBank. (WYMAN et al., 2004).

2.3.3.3 CEGMA

A ferramenta *Core Eukaryotic Genes Mapping Approach*, escrita na linguagem Perl, realiza a anotação automática de genomas eucarióticos utilizando modelos ocultos de Markov e conceitos de ortologia. Conta também com o algoritmo Fathom que calcula a acurácia dos éxons preditos pelos programas GeneWise e GeneId. Seus processos incluem múltiplos alinhamentos, uso do BLAST nas buscas no banco de dados KOG usando o alinhamento T-cofee para selecionar as proteínas de cada organismo. Um auto-treinamento pode ser realizado para a obtenção de melhores resultados (PARRA et al., 2007).

2.3.3.4 MARKER2

O MARKER2 é uma ferramenta escrita em Perl para gerenciamento e anotação de genomas sequenciados com a tecnologia de segunda geração. Utiliza processamento paralelo (multi-threads) para obter ganho de velocidade na execução dos processos.

Conta com os preditores de genes SNAP, Augustus e GeneMark (os dois últimos devem ser instalados e configurados pelo usuário), e com os buscadores InterProScan e BLASTx, este usado para identificar domínios no banco Pfam, representados por multiplos alinhamentos e modelos ocultos de Markov, além de dados providos pelo banco SwissProt (opcional). Sua anotação também pode contar com o auxilio de sequências de RNA-seq incorporadas no processo. Seu pacote está disponível para os sistemas Unix, Linux e MacOS (HOLT; YANDELL, 2011).

3 MÉTODOS

Levando em conta a abordagem comum básica para a anotação automática de genomas eucarióticos, ou seja, a estratégia que utiliza preditores gênicos (GOEL et al., 2013; LOH et al. 2015; MAKAROV, 2002; MATHÉ et al., 2002; WANG et al., 2004) para posteriormente tentar obter confirmações evidenciais da existência desses genes preditos, através de buscas por similaridades de sequências em banco de dados, explicaremos nessa seção a abordagem utilizada para o desenvolvimento do Sila Eukaryotic, o qual foi escrito em Matlab, arquitetado para plataforma Linux e seu funcionamento depende do pacote MCR (*Matlab Compiler Runtime*), disponível gratuitamente.

3.1 PREDIÇÃO DE GENES

Optamos por utilizar o programa GeneMark-ES (LOMSADZE et al., 2014) por sua quantidade de citações em artigos científicos, tipos de genomas testados, facilidade de instalação e execução, além dele necessitar apenas do arquivo de entrada contendo o genoma a ser analisado, sendo opcional a inserção de um arquivo de evidências EST.

Também tentamos inserir o preditor Genezilla (MAJOROS et al., 2005) no processo, pela sua quantidade de citações e testes publicados em artigos, porém tivemos problemas na compilação de uma biblioteca chamada “LBoom”, persistindo mesmo após seguir minuciosamente todas as instruções e recomendações descritas em seu manual e website, além de tentativas de contato por e-mail e em seu site oficial.

Visto que um dos objetivos deste projeto é melhorar a anotação a partir de uma predição, definiu-se que o GeneMark-ES seria um candidato suficiente, pois atendeu aos requisitos para integrar a ferramenta Sila Eukaryotic, ou seja, o arquivo de saída desse preditor contém as coordenadas dos genes e seus respectivos éxons.

3.2 BUSCAS POR SIMILARIDADE

Para a eleger o buscador de similaridade, optamos por testar e comparar o BLAT (KENT, 2002), devido ao alto desempenho em comparação ao BLAST (ALTSCHUL et al.,

1990), e o RAFTS3 (VIALLE et al., 2016), utilizando um computador com processador Intel Core-i5 (3.2 Ghz), 16 gigabytes de memória RAM e sistema operacional Ubuntu-BioLinux 8, com objetivo de determinar a sensibilidade, velocidade e consumo de recursos de hardware.

Extraímos aleatoriamente dez mil proteínas do banco GenBank-NR (2014) para compor a *query* (arquivo de consulta) de entrada, em formato multi-fasta. Posteriormente executamos as buscas por similaridades pelo BLAT e RAFTS3.

3.2.1 Bancos de dados

Para testar os buscadores de similaridade, utilizamos como argumento de entrada os bancos GenBank-NR (2014) e SwissProt (2015), ambos adquiridos do NCBI. Dividimos o banco de dados GenBank-NR em partes menores para obter uma estatística mais detalhada do comportamento dos dois buscadores, para então registrarmos seus ressultados e respectivos tempos de execução.

3.3 SEQUÊNCIAS GENÔMICAS

Selecionamos para análises as sequências de um cromossomo aleatório para quatro organismos diferentes, todos obtidos através do site do NCBI em março de 2015, sendo eles: *Aspergillus niger* (cromossomo 1), *Drosophyla melanogaster* (cromossomo 2), *Homo sapiens* (cromossomo 10) e *Oryza sativa* (cromossomo 1).

3.4 ANOTADORES

Para comparar a anotação do Sila Eukaryotic com outro anotador optamos por utilizar o MARKER2 (HOLT e YANDELL, 2011), devido ao seu ano de desenvolvimento mais atual e possibilitanto gerar melhores comparativos devido aos avanços e atualizações das ferramentas envolvidas em sua execução.

4 RESULTADOS E DISCUSSÃO

4.1 BUSCAS POR SIMILARIDADE

É importante ressaltar que, devido ao grande tamanho, dividimos o banco de dados GenBank-NR em pequenas partes para se obter uma estatística mais detalhada do comportamento dos dois buscadores (RAFTS3 e BLAT), então registramos seus respectivos tempos de execução.

Considerando que o BLAT faz a indexação no momento de sua execução, e não encontrou-se em seu manual uma opção de utilizar um banco já indexado, também medimos o tempo de indexação feita pelo RAFTS3, este que possibilita sua reutilização do banco (DBSTRUCT) já indexado anteriormente. Abaixo, a TABELA 3 mostra os tempos obtidos na execução dos testes de desempenho.

TABELA 3 – COMPARAÇÃO DE TEMPO ENTRE O BLAT E RAFTS3

GB-NR (2014)	Fastas (aa)	Tamanho (Gb)	Tempo em minutos com query de 10K fastas		
			BLAT	RAFTS	DBSTRUCT
100,00%	54.669.564	32,6	-	8,99	-*
23,00%	12.574.000	4,5	-	8,18	385,75
22,00%	12.027.304	4,3	67,38	8,13	365,22
20,00%	10.933.913	3,9	63,67	7,96	328,46
17,00%	9.293.826	3,3	58,51	7,77	283,57
12,00%	6.560.348	2,4	46,65	7,47	200,37
9,00%	4.920.261	1,7	37,36	6,04	152,04
6,00%	3.280.174	1,1	24,32	5,98	95,21
3,00%	1.640.087	0,5	12,91	5,90	51,04
1,50%	820,042	0,2	6,05	5,70	25,27

* Não medido anteriormente.
FONTE: O Autor (2015).

O algoritmo BLAT, durante o teste que utilizou o banco de referência com 23% do seu tamanho original, teve seu processo abortado/eliminado pelo sistema operacional em virtude do consumo de toda memória disponível no computador, incluindo a memória virtual (swap) de 16 GBytes.

Comparamos então a melhor proteína encontrada entre os buscadores, através de escore de similaridade, no teste que utilizou o GenBank-NR com 22% de seu tamanho original. O gráfico da FIGURA 11 mostra a comparação de similaridades obtidas pelo BLAT e RAFTS3.

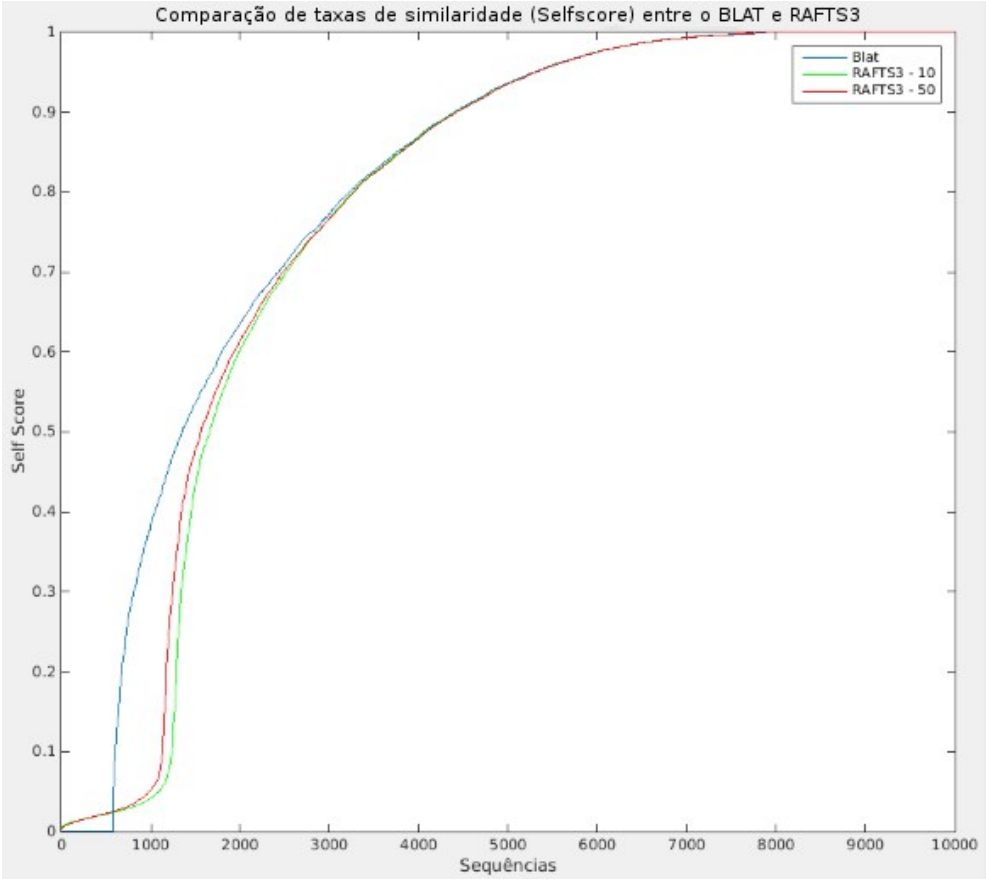

FIGURA11– COMPARAÇÃO DA SENSIBILIDADE ENTRE BUSCADORES TESTADOS
FONTE: O Autor (2015).

Apesar do BLAT não ter encontrado proteínas com similaridade para toda a query, ele apresentou maior sensibilidade para as proteínas com baixa similaridade em relação ao RAFTS3. Optamos então por aumentar o tamanho das sementes (k-mers) do RAFTS3, de 10 para 50 aminoácidos, para verificar seu ganho de sensibilidade. Como o ganho de sensibilidade, através dessa modificação, não mostrou-se superior ao do BLAT, decidimos manter as sementes em seu tamanho original (10 aminoácidos) em virtude do aumento do tempo de processamento versus o ganho de sensibilidade do RAFTS3.

Por fim, elegemos o comparador RAFTS3 pelos resultados mostrarem uma sensibilidade próxima a do BLAT, somando-se ao tempo e consumo de hardware entre as ferramentas, notando-se menos exigentes pelo algoritmo do RAFTS3.

4.2 SILA EUKARYOTIC

Para ser executado é necessário informar alguns parâmetros de utilização. Indica-se o caminho e nome do arquivo que contém o genoma a ser analisado e anotado. Como existe a opção de utilizar mais de um banco de dados de referência (GenBank-NR, SwissProt ou outro que possa ser indexado no formato do RAFTS3-DBSTRUCT), o usuário também deve indicar o caminho e nome do arquivo que contenha a estrutura de índices, e esse arquivo de índices já contém o nome do arquivo ao qual está indexado que é o próprio banco de dados. Caso o usuário não necessite de uma predição de genes, deverá ser informado também a não execução do preditor GeneMark-ES, determinado pelo argumento “0” (zero), ou “1” (um) para incluir o processo de predição. Se a opção de uso do preditor não for informado o Sila Eukaryotic executará os processos de predição somente caso seja necessário, ou seja, caso não exista informações de coordenadas de genes no arquivo de entrada (ex.: formato FASTA). Por fim, deve-se informar ao algoritmo o nome do arquivo GBK de saída, cujo conterá o genoma anotado no formato GenBank, exemplificado anteriormente na Figura 9. Se o nome do arquivo de saída não for informado, então o sistema irá criar um nome padrão com o nome formado por data e hora de sua execução. A linha de comando para chamar a execução do Sila Eukaryotic deverá ficar semelhante e com a seguinte forma: ~$ *./silae.sh <arquivo_entrada> <arquivo_dbstruct> <predição: 0 ou 1> <nome_arquivo_saída>* .

O Sila Eukaryotic inicia sua tarefa com os processos de leitura, checagem de extensão de arquivo e consistência dos dados nele contidos, além da organização das informações na memória RAM em estrutura de matriz, seja o arquivo no formato GBK ou FASTA. Então o próximo passo será o encaminhamento dos dados do genoma, ainda sem marcação de ORFs caso não existam, para o preditor de genes eucarióticos. A FIGURA 12 mostra o fluxo de dados do algoritmo no nível macro de visão.

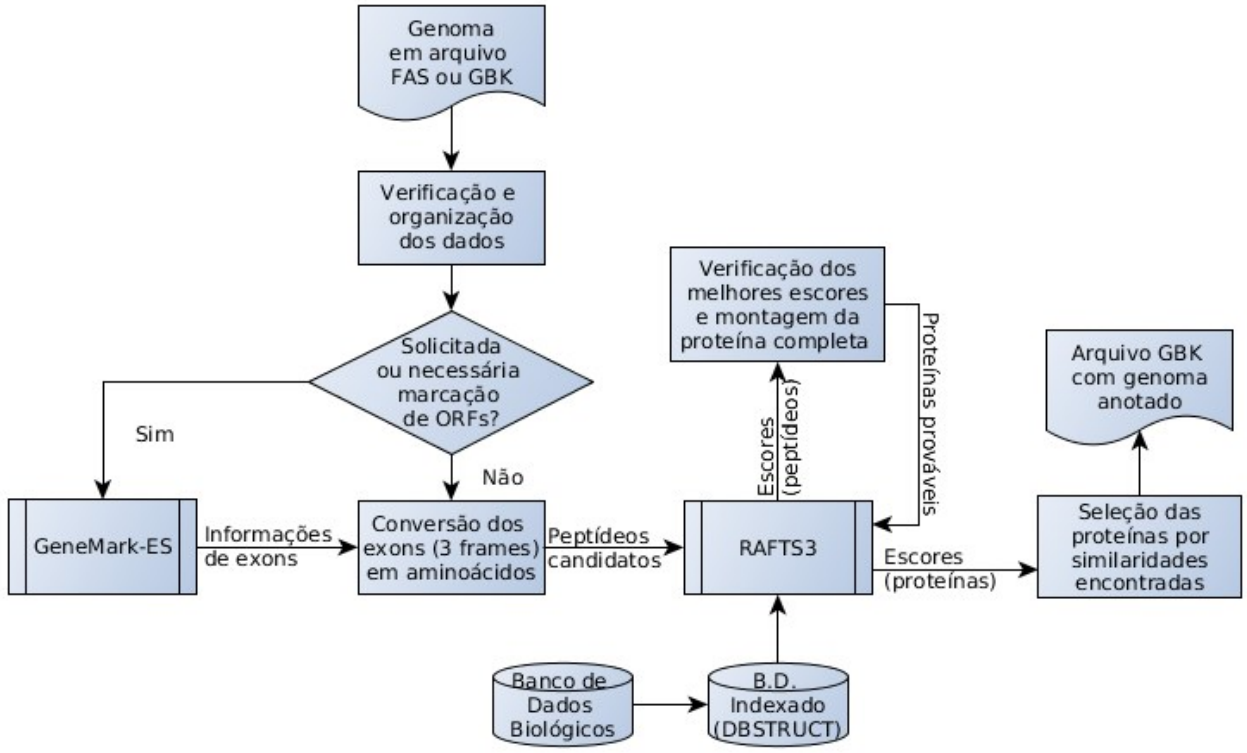

FIGURA 12 – FLUXOGRAMA DO SILA EUKARYOTIC
FONTE: O Autor.

Para a tarefa de predição de genes eucarióticos, opcional ao usuário, utilizamos o preditor GeneMark-ES, sendo que este faz parte do pacote Sila Eukaryotic como opção ao usuário. Caso o usuário já possua uma predição ou marcação gênica no formado GBK e não solicite nova predição, o sistema pulará esse estágio em que os éxons são analisados e preditos.

Considerando uma execução normal ou total do algoritmo, após a leitura e checagem do arquivo de entrada que contenha um genoma no formato FASTA ou GBK, as informações da sequência nucleotídicas serão repassadas para o GeneMark-ES realizar a predição de genes. Esse processo resultará num novo arquivo contendo coordenadas de exons, CDS, sentido da leitura nucleotídica (fitas complementares são lidas no sentido oposto), entre outros dados. Os resultados da predição serão organizados em matrizes, na memória RAM do computador, e indexados de modo a compor a matriz principal do sistema.

Todo exon predito será convertido para cada uma das três frames possíveis de leituras, calculadas com as variações dos tamanhos de códons (resultantes dos processos de *splicing* de maturação do mRNA), também verificando-se ausências de *stop*

codons, para sequências de aminoácidos. É necessário calcular a quantidade de nucleotídeos de cada éxon a fim de obter-se a sequência exata de aminoácidos, uma vez que o éxon pode ser cortado em regiões que dependerá do(s) nucleotídeo(s) contido(s) no próximo éxon. Nesta etapa, se o arquivo de entrada, no formato GBK, não necessitar de predição (e não for solicitada a predição padrão do Sila Eukaryotic), a leitura desse arquivo coletará informações como coordenadas de éxons, sentidos e demais, que passarão posteriormente pelo mesmo procedimento que os resultados do GeneMark-ES.

As sequências peptídicas possíveis, com no mínimo 30 aminoácidos, formarão a estrutura de query para primeira consulta de similaridade no banco de dados em uso, devendo este já estar no formato DBSTRUCT, ou seja, indexado para o RAFTS3 através do algoritmo “makeRafts3db” do seu próprio pacote. Contudo é válido lembrar que o pacote do Sila Eukaryotic conta com as opções de dois bancos já indexados em 2016: o GenBank-NR e o SwissProt. Essa query conterá as sequências individuais de cada éxon predito, e formará uma lista de peptídeos a fim de obter-se informações de similaridades com sequências de proteínas contidas no banco em uso. Essa estratégia tem como objetivo verificar a possibilidade erros ou equívocos de frames dos éxons, geralmente ocasionadas pelo fato do preditor considerar eventos padrões de splicing e não caber a um preditor verificar a evidência de similaridade em proteínas do éxon predito.

Continuando o processo do Sila Eucaryotic, os resultados de similaridade providos pelo RAFTS3, originados pela query de éxons, serão avaliados selecionando-se os melhores candidatos através dos escores de frames possíveis de éxons, também desconsiderando frames de éxons hipotéticos que possuam códons de parada (*stop codons*).

Então o algoritmo do Sila Eukaryotic montará as proteínas completas, concatenando os éxons com melhor escore das frames avaliadas. Caso a difereça de frame resulte num SelfScore inferior ao do preditor, este será descartado e substituído pelo éxon da frame indicada pelo preditor.

Novamente o RAFTS3 será executado, dessa vez para obtenção e identificação de proteínas com melhor similaridade (SelfScore), mas nessa consulta com a *query* contendo as proteínas sugeridas tanto pelo preditor quanto pelas melhores proteínas analisadas pelo Sila Eukaryotic. É importante esclarecer que o Sila Eukaryotic necessita de uma primeira consulta ao RAFTS3 para verificar todas as possibilidades de exons em relação às suas frames possíveis, e os escores de similaridades desses éxons podem possibilitar identificação de casos de “splicings alternativos”, estes não considerados nesse projeto,

mas passível de incrementações futuras que possibilitem facilmente essas identificações de “splicings alternativos”.

Após o processamento da etapa de segunda consulta, obtêm-se os novos resultados do RAFTS3. Serão selecionados os melhores escores de similaridade de cada gene finalista e então o programa criará o arquivo de saída, em formato GBK, contendo a anotação gênica, com uma anotação melhorada a partir de um preditor de genes eucarióticos. Também será criado um arquivo de LOG contendo informações de cada gene comparado, tais como: escores de similaridade, melhor anotação encontrada, etc. O nome do arquivo de LOG terá o mesmo nome do arquivo GBK, porém seu formato será com dados separados por tabulação, padrão texto ou “TXT”.

Para o arquivo de saída optamos por sinalizar os genes com a predição original do preditor, seja o GeneMark-ES ou outro externo utilizado pelo usuário. Abaixo, a FIGURA 13 mostra um exemplo de visualização, utilizando a ferramenta Artemis, dos resultados de uma anotação melhorada através da melhor similaridade, esta selecionada pelos melhores escores de cada proteína possível (hipotética ou conhecida), providos pelos processos e resultados da ferramenta de anotação automática Sila Eukaryotic.

FIGURA 13 – VISUALIZAÇÃO DO ARQUIVO DE SAÍDA DO SILA EUKARYOTIC

Tela do programa para visualização do arquivo GBK pelo programa Artemis.

FONTE: Software Artemis.

Através da diferenciação de cores, sendo a cor verde para os genes sugeridos pelo

Sila Eukaryotic e a cor vermelha para os genes sugeridos pelo preditor em uso (ou outro escolhido pelo usuário), o pesquisador poderá comparar mais facilmente as anotações feitas pela ferramenta Sila Eukaryotic com outra ferramenta qualquer.

4.2.1 Testes dos genomas

Após a execução do Sila Eukaryotic para cada um dos quatro organismos testados, capturamos todos os dados de genes e seus respectivos escores de similaridade obtidos dos dois bancos testados. Com o objetivo de comparar os escores e mensurar o resultado do Sila Eukaryotic, esses valores foram ordenados em ordem crescente, gerando então gráficos de curvas de desempenho atingido. As figuras 14, 15, 16 e 17, comentadas na sequência, mostram a comparação dos resultados dos escores obtidos com o preditor GeneMark-ES e com a ferramenta objeto deste trabalho (Sila Eukaryotic).

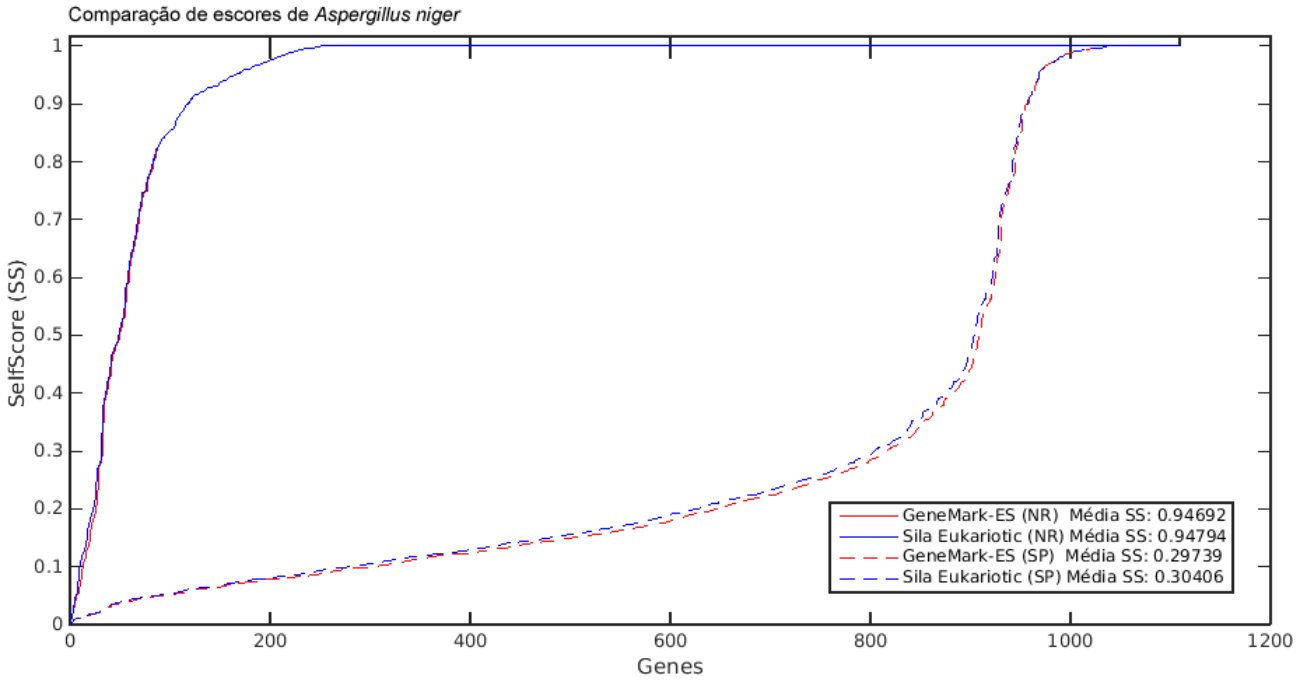

FIGURA 14 – Curva comparativa de escores dos bancos GenBank-NR e SwissProt.
As linhas contínuas representam os escores obtidos com o banco GenBank-NR e as linhas tracejadas com o banco SwissProt. São mostrados os resultados individuais do preditor e do anotador Sila Eukaryotic, pois este age somente a partir de uma predição ou anotação existente.
FONTE: O Autor (2016).

Podemos observar na FIGURA 14, que os escores obtidos com o banco NR foram

muito maiores que os obtidos com o banco SwissProt, isso devido ao RAFTS3 ter encontrado muitas proteínas hipotéticas de *Aspergillus niger* depositadas. Já com o banco SwissProt, que contém somente proteínas curadas, a média dos escores foi menor, com uma minoria de genes que atingiram 100% de similaridade. Vale acrescentar que quanto mais dados o banco de referência tiver, é muito provável que melhores escores de proteínas sejam encontrados.

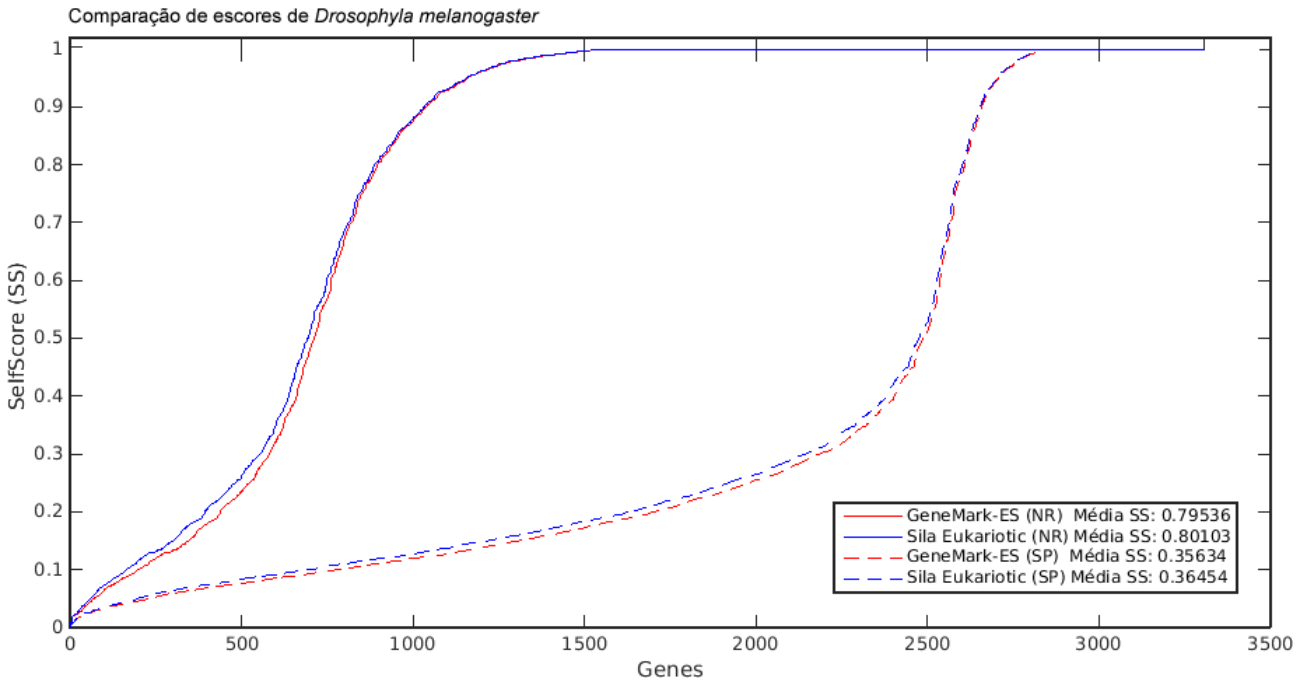

FIGURA 15 – Curva comparativa de escores dos bancos GenBank-NR e SwissProt.
As linhas contínuas representam os escores obtidos com o banco GenBank-NR e as linhas tracejadas com o banco SwissProt. São mostrados os resultados individuais do preditor e do anotador Sila Eukaryotic, pois este age somente a partir de uma predição ou anotação existente.
FONTE: O Autor (2016).

Na FIGURA 15 podemos observar um resultado semelhante ao obtido com o organismo *Aspergillus niger*, contudo o Sila Eukariotic conseguiu encontrar mais frames de éxons que possuem maior similaridade tanto com proteínas hipotéticas quanto com proteínas curadas, em relação ao preditor, mesmo que esses genes ainda possuam relativo baixo escore, 50% de similaridade ou menos, com as proteínas dos bancos.

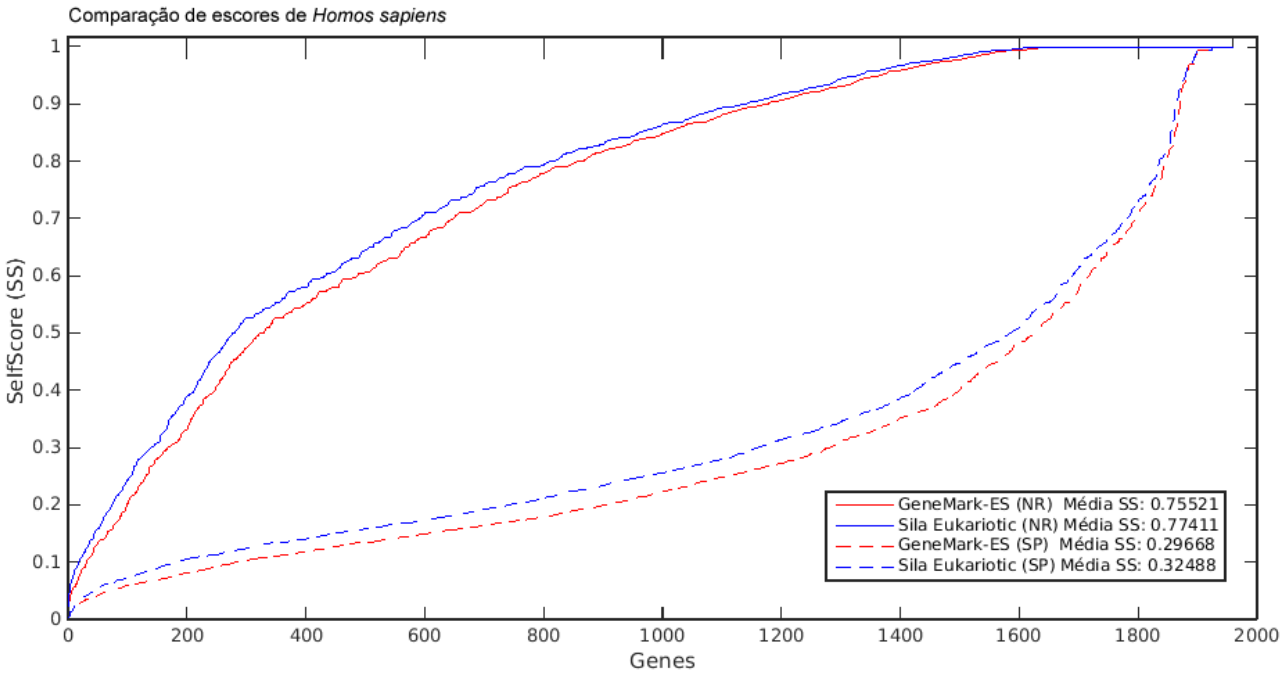

FIGURA 16 – Curva comparativa de escores dos bancos GenBank-NR e SwissProt.
As linhas contínuas representam os escores obtidos com o banco GenBank-NR e as linhas tracejadas com o banco SwissProt. São mostrados os resultados individuais do preditor e do anotador Sila Eukaryotic, pois este age somente a partir de uma predição ou anotação existente.
FONTE: O Autor (2016).

Para o organismo *Homo sapiens*, o Sila Eukaryotic obteve um resultado com melhora mais evidente na FIGURA 16, se comparado aos dois organismos anteriores. Porém, seguiu o mesmo padrão para os genes de alta similaridade, ou seja, os resultados entre o algoritmo e o preditor utilizado, independente do banco de referência.

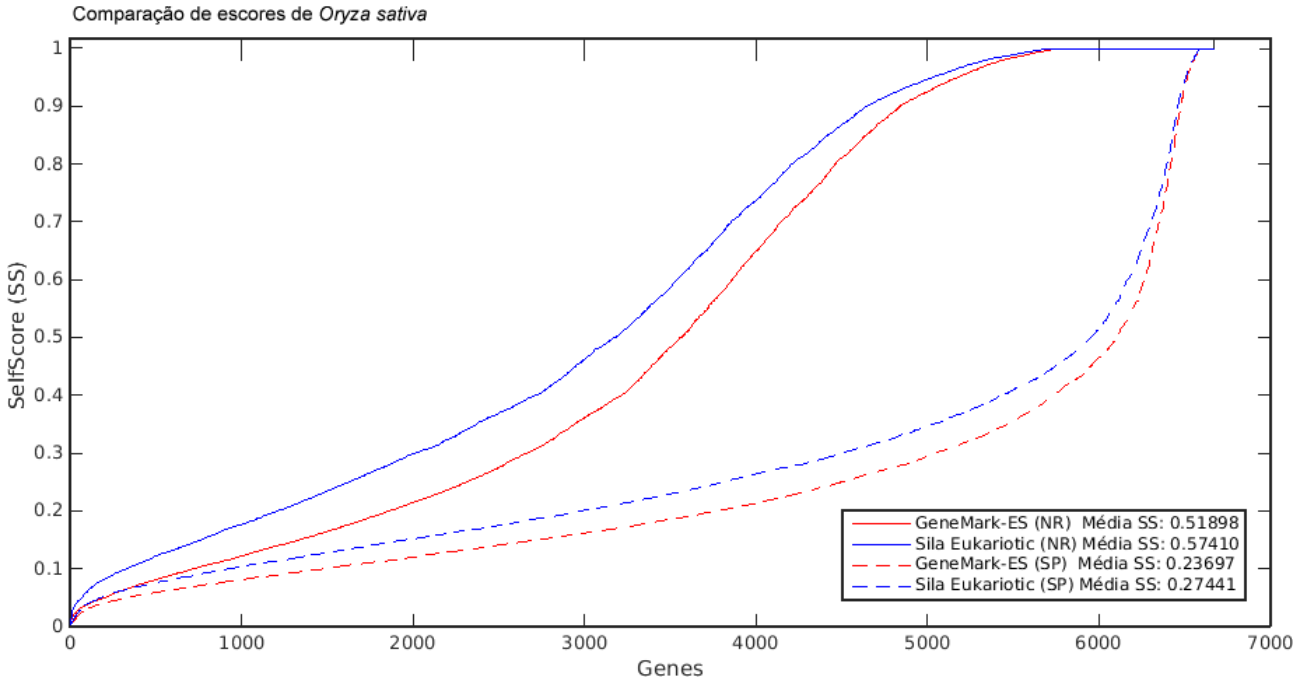

FIGURA 17 – Curva comparativa de escores dos bancos GenBank-NR e SwissProt.
As linhas contínuas representam os escores obtidos com o banco GenBank-NR e as linhas tracejadas com o banco SwissProt. São mostrados os resultados individuais do preditor e do anotador Sila Eukaryotic, pois este age somente a partir de uma predição ou anotação existente.
FONTE: O Autor (2016).

Entre os quatro organismos testados, a sequência de *Orysa sativa* foi em que obteve-se a mais alta melhora de escores pela seleção de frames dos éxons candidatos, gerando uma proteína com maior similaridade em relação ao predito pelo GeneMark-ES.

Calculamos então a melhora de escores através das médias dos escores obtidos, tanto do GeneMark como do Sila Eukaryotic, para obter uma visão geral dos resultados. A TABELA 4 mostra, além das médias de score de cada organismo testado, a porcentagem de escore de similaridade obtida com os dois bancos de referência utilizados.

TABELA 4 – COMPARAÇÃO DA MÉDIA DOS ESCORES DE SIMILARIDADES

		Médias de Similaridade (SelfScore*)				Melhora de Escores	
		GeneMark-ES		Sila Eukaryotic			
Organismo	Cromossomo	SwissProt	GenBank(NR)	SwissProt	GenBank(NR)	SwissProt	GenBank(NR)
Oryza sativa	1	0,23697	0,51898	0,27441	0,57410	3,74%	5,51%
Homo sapiens	10	0,29668	0,75521	0,32448	0,77411	2,78%	1,89%
Drosophylla melanogaster	2	0,35634	0,79536	0,36454	0,80103	0,82%	0,57%
Aspergillus niger	2	0,29739	0,94692	0,30406	0,94794	0,67%	0,10%

SelfScore é uma medida de varia de 0 (zero: similaridade inexistente) até 1 (um: totalmente similar ou idêntico).São mostrados os resultados individuais do preditor e do anotador Sila Eukaryotic em virtude deste funcionar somente a partir de uma predição ou anotação já existente.
FONTE: O Autor (2016).

É possível notar na tabela o aumento geral de escores de similaridades entre genes e proteínas dos bancos, sendo o menor ganho (0,1%) obtido para o organismo *Aspergillus niger,* utilizando o banco NR . O maior aumento foi obtido também com o NR mas para o organismo *Oryza sativa* (5,51%), talvez por alguma particularidade do preditor GeneMark-ES. Apesar da média de escores obtidas com o banco SwissProt ser menor, o Sila Eukaryotic resultou num maior aumento de similaridades encontradas, para três dos quatro organismos, comparadas ao banco NR.

É importante destacar que o algoritmo do Sila Eukaryotic não poderá gerar perdas nos escores, uma vez que ele sempre optará pela melhor anotação, ou seja, a que tiver maior escore, então na pior possibilidade a melhora será igual a zero.

4.3 COMPARAÇÃO ENTRE ANOTADORES

Para comparar o algoritmos anotadores MARKER2 e Sila Eukaryotic, utilizamos a sequência de *Homo sapiens* como argumento de entrada. Elencamos características de ambos os softwares e por fim registramos seus resultados. A TABELA 5 elenca as principais características entre eles.

TABELA 5 – COMPARAÇÃO DE CARACTERÍSTICAS ENTRE ANOTADORES

	Algoritmos anotadores	
Características	**Sila Eukaryotic**	**MAKER2**
Dependências externas	MCR	Python e Perl
Preditor embutido	GeneMark-ES	SNAP
Buscador de Similaridades	Rafts3	RepeatMasker e BLASTx
Interface	Linha de comando	Linha de comando
Linguagem	Matlab	Perl
Banco de dados compatível	DBStruct*	Multi-Fasta
Formato de entrada	Fasta e GBK	Fasta
Formato de saída	GBK	GFF3
Aceita arquivo já anotado	Sim	Não
Marcação de exons	Sim	Não**
Anotação de genes	Sim	Sim
Geração de LOG para apoio	Sim	Sim
Suporte nativo de *multithread*	Não	Sim
Sistema Operacional	Linux	Linux e MacOS (Darwin)

A tabela destaca as principais características constatadas entre os softwares anotadores.

(*) Formato específico de indexação de banco de dados multi-fasta para o Rafts3.

(**) Alguns exons são marcados quando i arquivo EST de mRNA.

FONTE: O Autor (2016).

A instalação do MARKER2 exigiu o cadastro no site *www.girinst.org*, que mantém um banco de dados chamado RepBase, utilizado pelo algoritmo RepeatMasker que vem incorporado no pacote de instalação. Após o cadastro deve-se aguardar um e-mail de contato, cujo e-mail solicitante não pode ser comercial (não possuir ".com" no endereço), dispobilizando assim o link para baixar o RepBase. Para a execução padrão do MARKER2, não é obrigatório informar um banco de dados de referência de proteínas, porém os demais algoritmos ficam fora do processo, o que resulta num arquivo sem regiões identificadas, apenas com marcações de regiões repetidas, que podem ser possíveis genes. Para o usuário gerar um arquivo devidamente anotado, é necessário, assim como no Sila Eukaryotic, informar um banco de dados de proteínas. É importante também informar um arquivo de EST (evidências de sequências mitocondriais) para então o resultado ser mais completo, com identificação das regiões de éxons, mesmo que alguns genes possam ficar sem as coordenadas de seus éxons. Diferentemente do Sila Eukaryotic, o MARKER2 necessida que seus argumentos sejam salvos em arquivos de configurações chamados de "controles". Esses arquivos guardam os caminhos dos demais algoritmos envolvidos (tais como genoma de entrada, EST, preditor, buscador, etc), e após informar os caminhos do banco de dado e arquivo de genoma a ser analisado o anotador estará pronto para ser executado.

Realizamos os testes utilizando o mesmo banco de dados, porém não passando um arquivo EST para o MARKER2 e utilizando sua execução padrão, ou seja, sem alterar nenhum dado de configuração de nenhum anotador. A TABELA 6 mostra os resultados obtidos.

TABELA 6 – RESULTADOS DOS ANOTADORES

	Sila Eukaryotic	**MAKER2**
Configurações	Padrão	Padrão
Banco de dados *	SwissProt	SwissProt
Formado to arquivo de entrada	Fasta	Fasta
Tempo de execução (minutos)	103,1	2335,3
Regiões de genes encontradas	1857	1364**
Regiões de *exons* encontradas	14619	-***
Tamanho total do output (Mbytes)	51,99	222,1

A tabela mostra os resultados dos testes feitos no mesmo computador, utilizando a sequência do cromossomo 10 do organismo *Homo sapiens*.

(*) Versão obtida no NCBI em Maio/2015. (**) Valor obtido a partir das coordenadas de genes que não se sobrepuseram. (***) Arquivos de saída não contêm coordenadas de exons.

FONTE: O Autor (2016).

Considerando que o MARKER2 não informou as coordenadas de todos os éxons, não foi possível calcular o escore de similaridade entre proteínas para ele, visto que em genomas eucarióticos a proteína geralmente é formada por um conjunto de éxons. Consequentemente não foi possível realizar um teste para verificar se o Sila Eukaryotic melhoraria a anotação do MARKER2, convertendo o formato GFF3 para GBK, em virtude de não haver quantidade e informações de éxons suficientemente ao Sila Eukaryotic.

O tamanho em bytes da pasta de resultados do MARKER2 é maior, pelo registro de vários eventos em arquivos de log, bem como cópia dos seus arquivos de controle e a própria anotação em formato GFF3, contendo muitas coordenadas redundantes do conjunto de algoritmos envolvidos, entre eles o BLASTx, RepeatMasker e Protein2genome.

5 CONCLUSÃO

É importante lembrar que o Sila Eukaryotic estatisticamente proporcionará melhores resultados conforme o acrécimo de sequências no banco de dados utilizado, ainda que possa aumentar o tempo necessário para sua execução. A FIGURA 18 exemplifica e mostra a comparação do aumento de similaridades encontradas com a atualização do banco GenBank-NR.

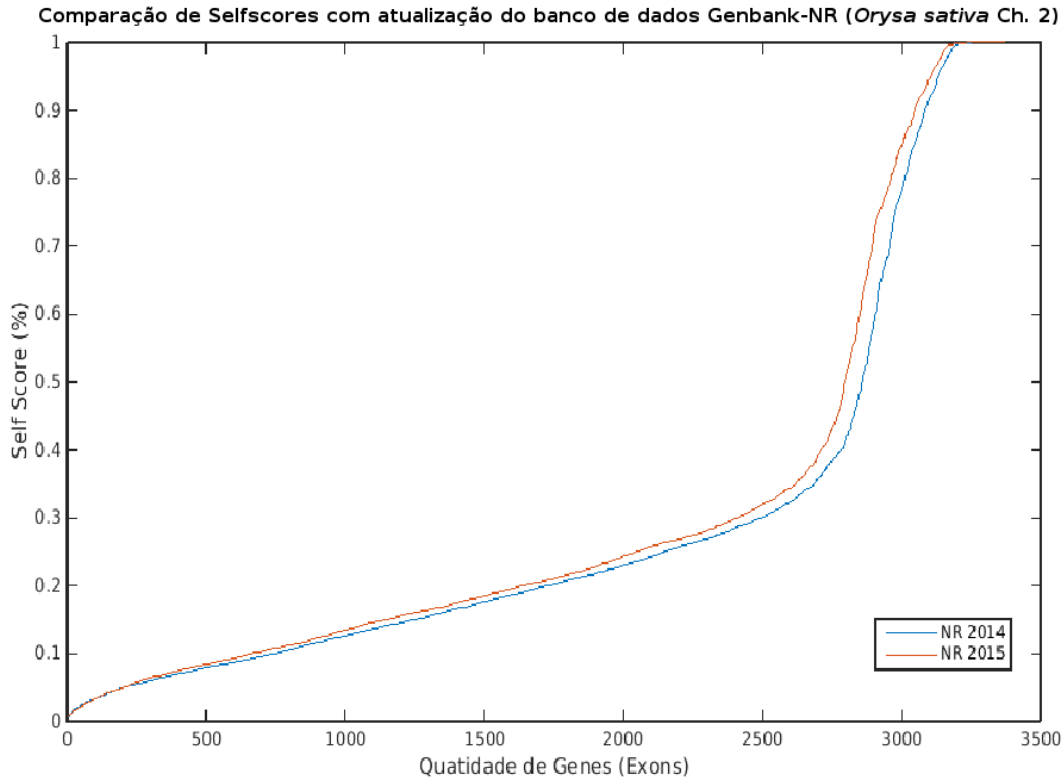

FIGURA 18 – COMPARAÇÃO DE TAXAS DE SIMILARIDADES OBTIDAS COM O RAFTS3 APÓS ATUALIZAÇÃO DO BANCO DE DADOS
FONTE: O Autor (2015).

Isso significa que, com o passar do tempo melhor será o resultado do Sila Eukaryotic, pois o depósito de novas proteínas (hipotéticas ou curadas) nos bancos proporcionará mais sequências para serem comparadas.

Desenvolvemos neste trabalho uma ferramenta que faz a anotação automática de genomas eucarióticos com alto desempenho de velocidade e que não exige grandes recursos de hardware. O Sila Eukaryotic mostrou-se eficiente na melhora da anotação através de uma predição e medidas de similaridades de proteínas.

O pacote do sistema de anotação está disponível para execução local no site www.sourceforge.com, e oferece facilidade na instalação e utilização, também proporcionando rapidez nos resultados para usuários que necessitem da anotação genômica eucariótica.

Concluímos então que este trabalho apresenta uma ferramenta alternativa e rápida para anotação automática de organismos eucarióticos, proporcionando melhora na anotação a patir de uma predição de genes.

6 RECOMENDAÇÕES E PROJETOS FUTUROS

Como o Sila Eukaryotic foi escrito na linguagem Matlab, que depende de um interpretador chamado de MCR, e mesmo o algoritmo podendo ser compilado para um executável, é possível que sua reescrita na linguagem C/C++ dê ao programa mais velocidade de execução.

O acrécimo de outros preditores no Sila Eukaryotic poderá melhorar seus resultados finais de anotação, visto que preditores podem identificar regiões gênicas diferentes num mesmo genoma e anotadores podem utilizar mais de um meio para predição. Isso significa que havendo mais regiões a se considerar, haverão mais possibilidades de melhora na anotação.

A disponibilização da ferramenta Sila Eukaryotic em plataforma Web também poderá ocasionar melhoras , simplificando sua utilização além de não ser necessário a instalação locl sua utilização e sem a necessidade de instaalar num computados local.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ALLEN, J. E.; PERTEA, M.; SALZBERG, S. L. Computational Gene Prediction Using Multiple Sources of Evidence. **Genome Research**, v. 14, p. 142-148, 2004.

ALLEN, J. E.; SALZBERG, S. L. JIGSAW: integration of multiple sources of evidence for gene prediction. **Bioinformatics**, v. 21, n. 18, p. 3596-3603, 2 ago. 2005.

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403-410, 5 out. 1990.

BAIROCH, A.; APWEILER R. The Swiss-Prot protein sequence data bank and its supplement TrEMBL. **Nucleic Acids Research**, n. 25, p. 31-36, 1997.

BAKKE, P. et al. Evaluation of three automated genome annotations for Halorhabdus utahensis. **PloS one**, v. 4, n. 7, p. 6291, jan. 2009.

BAXEVANIS, A.; OUELLETTE, B. **Bioinformatics**. New York, USA: John Wiley & Sons, Inc., v. 43. 2001.

BERG, J. M.; TYMOCZKO, J. L.; STRYER, L. **Bioquímica**. 6. ed. Rio de Janeiro: Guanabara Koogan, 2010.

BERRIS, G. F.; KING, O. D.; BRYANT, B.; SANDER, C.; ROTH, F. P. Characterizing gene sets with FunAssociate. **Bioinformatics Applications Note**, v. 19, n. 8, p. 2502-2504, abr. 2002.

BLATTNER, F. R. et al. The complete genome sequence of Escherichia coli K-12. **Science** (New York, N.Y.), v. 277, n. 5331, p. 1453-1462, 5 set. 1997.

BURGE, C.; KARLIN, S. Prediction of Complete Genome Gene Structures in Human Genomic DNA. **JMB**, v. 268, p. 78-94, 1997.

CRICK, F. Central dogma of molecular biology. **Nature**, v. 277, p. 561-563, 8 ago. 1970.

FARLEY, D. W.; DONALDSON, S. L.; COMES, O.; ZUBERI, K.; BADRAWI, R.; CHAO, P.; FRANZ, M.; GROUIOS, C.; KAZI, F.; LOPES, C. T.; et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. **Nucleic Acids Reserach**, v. 38, p. Web Server Issue, fev. 2010.

FLEISCHMANN, R. D. et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. **Science** (New York, N.Y.), v. 269, n. 5223, p. 496-512, 28 jul. 1995.

GOEL, N.; SINGH, S.; ASERI, T. C. A Review of Soft Computing Techniques for Gene Prediction. **ISRN Genomics**, v. 2013, 26 dez. 2012.

GUIGÓ, R.; KNUDSEN, S.; DRAKE, N.; SMITH, T. Prediction of Gene Structure. **Molecular Biology Computer Research Resource**, v. 226, p. 141-157, set. 1992.

HENDERSON, J.; SALZBERG, S.; FASMAN, K. H. Finding Genes in DNA with Hidden Markov Model. **Journal of Computacional Biology**, v. 4, n. 2, p. 127-141, 1997.

HOLT, C.; YANDELL, M. MARKER2: an annotation pipeline and genome-database management tool for second-generation genomes projects. **BMC Bioinformatics**, dez. 2011.

HOOPER, P. M.; ZHANG, H.; WISHART, D. S. Preditcion of genetic srtucture in eukaryotic DNA using reference point logistic regression and sequence alignment. **Bioinformatics**, v. 16, n. 5, p. 425-438, 2000.

HWANG, S; RHEE, S. Y.; MARCOTTE, E. M.; LEE, I. Systematic prediction of gene function in *Arabidopsis thaliana* using a probabilistic funcional gene network. **Nature America**, v. 6, n. 9, p. 1429-1442, 25 ago. 2011.

KENT, J. K. BLAT - The BLAST-Like Alignment Tool. **Genome Research**, v. 12, p. 656-664, mar. 2002.

KOONIN, E. V; GALPERIN, M. Y. **Sequence - Evolution - Function Computational Approaches in Comparative Genomics**. Boston: Kluwer Academic; 2003.

KROGH, A. Using Database Matches with HMMGene for Automated Gene Detection in Drosophila. **Genome Research**, n. 10, p. 523-528, 2000.

KULP, D.; HAUSSLER, D.; REESE, M. G.; EECKMAN, F. H. A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA. **ISMB**, v. 96, p. 134-142, 1996.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860-921, 15 fev. 2001.

LEWIS, S.; ASHBURNER, M.; REESE, M. G. Annotating eukaryote genomes. **Current opinion in structural biology**, v. 10, n. 3, p. 349-354, jun. 2000.

LIPMAN, D.; PEARSON, W. Rapid and sensitive protein similarity searches. **Science**, v. 227, n. 4693, p. 1435-1441, 22 mar. 1985.

LOH, S. K.; LOW, S. T.; MOHAMAD, M. S.; DERIS, S.; KASIM, S.; WEN, C. Y.; IBRAIM, Z.; SUSILO, B.; HENDRAWAN, Y.; WARDANI, A. K. A Review of Software for Predicting Gene Function. **International Journal of Bio-Science and Bio-Technology**, v. 7, n. 2, p. 50-70, 2015.

LOMSADZE, A.; TER-HOVHANNISYAN, V.; CHERNOFF, Y. O.; BORODOVSKY, M. Gene identification in novel eukaryotic genomes by self-training algorithm. **Nucleic Acids Research**, v. 33, n. 20, p. 6494-6506, 28 nov. 2005.

LUKASHIN, A. V.; BORODOVSKY, M. GeneMark.hmm: new solution for gene finding. **Nucleic Acids Research**, v. 26, n. 4, p. 1107-1115, ago. 1997.

MAJOROS, W. H.; PERTEA, M.; SALZBERG, S. L. TrigrScan and GlimmerHMM: two open source ab initio eykaryotic gene-finders. **Bioinformatics**, v. 20, n. 16, p. 2878-2879, 14 mai. 2004.

MAJOROS, W. H.; PERTEA, M.;SALZBERG, S. L. Eficient implementationof a generalized pair hidden Markov model for comparative gene finding. **Bioinformatics**, v. 21, n 9, p. 1782-1788, jan. 2005.

MAKAROV, V. Computer programs for eukaryotic gene prediction. **Briefings in Bioinformatics**, v. 3, n. 2, p. 195-199, jun. 2002.

MASSOUNI, N.; RIVERA, C. G.; MURALI, T. M. VIRGO: computational prediction of gene functions. **Nucleic Acids Research**, v. 34, n. Web server issue, p. 340-344, fev. 2006.

MATHÉ, C.; SAGOT, M.; SCHIEX, T.; ROUZÉ, P. Current methods of gene prediction, their strengths and weaknesses. **Nucleic Acids Research**, v. 30, n. 19, p. 4103-4117, mai. 2002.

MYERS, C. L.; ROBSON, D.; WIBLE, A.; HIBBS, M. A.; CHIRIAC, C.; THEESFELD, C. L.; DOLINSKI, K; TROYANSKAYA, O. L. Discovery of biological networks from diverse funcional genomic data. **Genome Biology**, v. 6, n. issue 13, p. 6:R114, dez. 2005.

NELSON, D. L.; COX, M. M. **Lehninger princípios de bioquímica**. 4. ed. São Paulo: Sarvier, 2006.

PARRA, G.; BRADMAN, K.; KORF, I. CEGMA: a pipeline to accuretly annotate core genes in eucaryotic genomes. **Bioinformatics**, v. 23, n. 9, p. 1061-1067, 7 dez. 2006.

PETTY, N. K. Genome annotation: man versus machine. **Nature reviews**. Microbiology, v. 8, n. 11, p. 762, nov. 2010.

PROSDOCIMI, F. Curso Online - INTRODUÇÃO A BIOINFORMÁTICA. 2007. Disponível em <http://www2.bioqmed.ufrj.br/prosdocimi/Fprosdocimi07_CursoBioinfo.pdf>. Acesso em 12 de janeiro de 2015.

QUAIL, M. A et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. **BMC genomics**, v. 13, n. 1, p. 341, jan. 2012.

ROUZÉ, P.; PAVY, N.; ROMBAUTS, S. Genome annotation: which tools do we have for it? **Current opinion in plant biology**, v. 2, n. 2, p. 90-5, abr. 1999.

SALAMOV, A. A.; SOLOVYEV, V. V. Ab initio Gene Finding in Drosophila Genomic DNA. **Genome Research**, v. 10, p. 516-522, 2000.

SALLET, E.; GOUZY, J.; BARDOU, P.; CROS, M.; FOISSAC, S.; MOISAN, A.; NOIROT, C.; LEROUX, D.; SCHIEX, T. EuGène: an open gene finder for eukaryotes and prokaryotes. **Applied Mathematics and Computer Science Dept.** INRA Toulouse-France, dez. 2014.

SALZBERG, S.; DELCHER, A. L.; FASMAN, K. H.; HENDERSON, J. A Decision Tree System dor Finding Genes in DNA. Johns Hopkins University, Baltimore. 1998.

SANGER, F.; NICKLEN, S. DNA sequencing with chain-terminating. **Proc. Natl. Acad. Sci. USA**, v. 74, n. 12, p. 5463-5467, dez. 1977.

SCHWEIKERT, G.; BEHR, J.; ZIEN, A.; ZELLER, G.; ONG, C. S.; SONNENBURG, S.; RATSCH, G. mGene.web: a web service for accurate computational gene finding. **Nucleic Acids Research**, v. 37, p. 312-316, jun. 2009.

SNYDER, E. E.; STORMO, G. D. Identification of Protein Coding Regions in Genomic DNA. **JMB-MS**, n. 435, p. 1-18, 1995.

SOLOVYEV, V.; KOSAREV, P.; SELEDSOV, I; VOROBYEV, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. **Genome Biology**, v. 7, p. S10.1-S10.12, 7 ago. 2006.

STEIN, L. Genome annotation: from sequence to biology. **Nature reviews genetics**, v. 2, p. 493-503, Jul, 2001.

STOESSER, G,. et al. The EMBL Nucleotide Sequence Database. **Nucleic Acids Research**, v. 30, p. 21-26, 2002.

VIALLE, R. A. SILA – Um Sistema para Anotação Automática de Genomas Utilizando Técnicas Independente de Alinhamento. Dissertação – Universidade Federal do Paraná. Curitiba. 2013.

VIALLE, R. A.; PEDROSA, F. O.; WEISS, V. A.; GUIZELINI, D.; TIBAES, J. H.; MARCHAUKOSKI, J. N.; SOUZA, E. M.; RAITTZ, R. T. RAFTS3: Rapid Alignment-Free Tool dor Sequence Similarity Search. **BioRxiv**, 31 mai. 2016.

WANG, Z.; CHEN, Y.; LI, Y. A Brief Review of Computational Gene Prediction Methods. **Geno. Prot. Bioinfo**., v. 2, n. 4, nov. 2004.

WATSON, J. D. et al. **Biologia molecular do gene**. 5. ed. Porto Alegre: Artmed, 2006. p. 760. 2006.

WIRTH, A. I. A Plasmodium falciparum Genefinder. Honours research project. Department of Mathematics and Statistics, University of Melbourne. 2001.

WYMAN, S. K.; JASEN, R. K.; BOORE, J. L. Automatic annotation of organellar genomes with DOGMA. **Bioinformatics**, v. 20, n. 17, p. 3252-3255, 4 jun. 2004.

XU, Y.; EINSTEIN, J. R.; MURAL, R. J.; SHAH, M.; UBERBACHER, E. C. An Improved System for Exon Recognition and Gene Modeling in Human DNA Sequences. **ISBM**, v. 94, 1994.

YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics**, v. 13, p. 329-342, 18 abr. 2012.

ZHANG, M. Q. Identification of protein coding regions in the human genome by quadratic discriminant analysis. **Genetics**, v. 94, p. 565-568, jan. 1997.