Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

Leishmania braziliensis: Reanotação genômica e estruturação das informações em modelos de dados relacionais

Felipe Guimarães Torres

Feira de Santana

2016

Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

Felipe Guimarães Torres

Leishmania braziliensis: Reanotação genômica e estruturação das informações em modelos de dados relacionais

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Orientador: Dr. Artur Trancoso Lopo de Queiróz

Coorientador: Dr. Vinicius Maracajá-Coutinho

Feira de Santana

2016

Felipe Guimarães Torres

Leishmania braziliensis: Reanotação genômica e estruturação das informações em modelos de dados relacionais

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Feira de Santana, 18 de dezembro de 2015

BANCA EXAMINADORA

Prof. Dr. Artur Trancoso Lopo de Queiróz (Orientador)
Centro de Pesquisas Gonçalo Moniz da Fundação Oswaldo Cruz

Prof. Dr. Iranderly Fernandes de Fernandes
Departamento de Física da Universidade Estadual de Feira de Santana

Prof. Dr. Antonio Ricardo Khouri Cunha
Centro de Pesquisas Gonçalo Moniz da Fundação Oswaldo Cruz

Ficha Catalográfica – Biblioteca Central Julieta Carteado

Torres, Felipe Guimarães
T818f *Leishmania braziliensis* : reanotação genômica e estruturação das informações em modelos de dados relacionais / Felipe Guimarães Torres. – Feira de Santana, 2016.
49 f. : il.

Orientador: Artur Trancoso Lopo de Queiróz.
Coorientador: Vinicius Maracajá-Coutinho.

Dissertação (mestrado) – Universidade Estadual de Feira de Santana, Programa de Pós-Graduação em Computação Aplicada, 2016.

1. Computação – Genética. 2. *Leishmania braziliensis*. – Genoma – Banco de dados. I. Queiróz, Artur Trancoso Lopo, orient. II. Maracajá-Coutinho, Vinicius, coorient.. III. Universidade Estadual de Feira de Santana. IV. Título.

CDU: 004.65:616.928.5

Abstract

Cutaneous leishmaniasis affects 12 million people around the world. The main etiological agent in Brazil is Leishmania braziliensis. Its annotation is not very well characterized and have inaccurate regions because of the large number of hypothetical and putative genes. To solve this problem, we annotated the Leishmania braziliensis genome (MHOM / BR / 75 / M2904) with new algorithm and curated database by similarity. Initially, we downloaded these sequences of Leishmania braziliensis from NCBI and used GENSCAN and GLIMMER algorithms. We compared the predicted genes with SWISSPROT protein database. The comparison process was made by BLASTx and SWISS-PROT. We also used StructRNAFinder to predict the non-coding RNA, ncrna. We identified 25,539 ORF's, 4,916 genes, 735 ncRNA and 863 proteins. 94.75% of predicted genes from Leishmania braziliensis were present on Leishmania panamensis genome (MHOM/PA/94/PSC-1). The comparison of ncrna number and proteins highlights a relation between chromosome and ncrna. All genes found were mapped and aligned to the Leishmania braziliensis genome and stored in a relational database model built in PHP 5.0 and MySQL. All data resulting of analyses in this work is available at www.leishdb.com.

Keywords: leishmania braziliensis, genomic annotation, bioinformatics, genome prediction, biological database

Resumo

A leishmaniose cutânea afeta cerca 12 milhões de pessoas ao redor do mundo. O principal agente etiológico dessa patologia no Brasil é a Leishmania braziliensis. A anotação dela não está bem caracterizada e possui regiões sem uma grande certeza devido ao grande número de genes hipotéticos e putativos. Para resolver esse problema, nós reanotamos o genoma da *Leishmania braziliensis* (MHOM/BR/75/M2904) com novas versões de algoritmos e uma base de dados curada. Inicialmente, nós baixamos e predizemos os genes com uma base de proteínas. A comparação foi feita por BLASTx e o banco de dados SWISS-PROT. Também utilizamos o preditor Struct RNA Finder para predição de RNA's não codificantes. Como resultado desse trabalho identificamos 25539 ORF's, 4916 genes, 735 ncRNA e 863 proteínas. Cerca de 94.75% genes preditos também estavam presentes em anotação da Leishmania panamensis (MHOM/PA/94/PSC-1). Comparando o número de rna não codificante e proteínas foi evidenciado a presença de relações entre cromossomos e ncrna. Todos os genes encontrados foram mapeados no genoma da Leishmania braziliensis e armazenados em um modelo de banco de dados relacional construído em MySQL e PHP 5.0. Todos os dados resultantes desse trabalho estão disponíveis no www.leishdb.com.

Palavras-chave: leishmania braziliensis, anotação genômica, bioinformática, predição gênica, banco de dados biológicos

Prefácio

Esta dissertação de mestrado foi submetida à Universidade Estadual de Feira de Santana (UEFS), como requisito parcial para obtenção do grau de Mestre em Computação Aplicada.

A dissertação foi desenvolvida dentro do Programa de Pós-Graduação em Computação Aplicada (PGCA), tendo como orientador o Dr. **Dr. Artur Trancoso Lopo de Queiróz** e co-orientador o Dr. **Vinicius Maracajá-Coutinho**.

Esta pesquisa foi financiada pela Fapesb.

Agradecimentos

Agradeço inicialmente a Deus por ter me ajudado até aqui e dado sabedoria para lidar com as adversidades.

Agradeço a minha namorada, Claudiana de Carvalho Melo que a todo momento esteve me apoiando e incentivando nos momentos mais difíceis, também sorrindo e comemorando os momentos de conquistas e felicidades.

Não poderia esquecer dos colegas e amigos: Adélia Icó, Leonardo Melo e José Irahe Kasprzykowski, que por muitas vezes me incentivaram e ajudaram durante toda a caminhada. Seria injusto se esquecesse dos colegas e amigos pela ajuda em momentos de dificuldade: Kiyoshi Fukutani, Camila Indiani, Ivonete Marciel e Ricardo Khouri. Meus sinceros agradecimentos ao Gabriel Santos, Raquel Santos, Aildes Matos que durante os momentos mais complicados tornaram esse sonho possível, obrigado a RH Soft por sempre apoiarem e serem pessoas de referência para a minha vida de integridade, amizade e lealdade acima de tudo. Agradeço também aos professores, verdadeiros mestres e agora doutores (rsrsrs): Márcio Soussa, Patrícia Braga e João Vasconcellos, pelos conselhos, incentivos e torcida por mais esta etapa da minha carreira acadêmica. A eles devo a minha iniciação na área acadêmica e aprendizados importantes que levarei por toda vida, sou eternamente grato a vocês por isso.

Gostaria de agradecer também a minha prima, Mariana Menezes Alencar por todos os conselhos, apoios e auxílios, obrigado por ser um anjo na minha vida e estar presente nela a cada momento. Agradeço também aos meus pais por terem me apoiado. Ao meu avó Gilvan de Menezes Guimarães, agradeço o brilhante exemplo e incentivo que a sua vida traduz a todos que a observam. Agradeço a minhas avós, meus tios e primos por estarem torcendo a cada momento.

Agradeço também a Vinícius Maracajá pela coorientação e todas as ajudas e conselhos nos momentos mais difíceis da longa jornada.

Agradeço por fim ao meu orientador e amigo o Prof. Dr. Artur Trancoso Lopo de Queiróz pelos conselhos, orientações e puxões de orelhas que me formaram como o profissional e acadêmico que sou hoje.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Abrebiações

Abreviação	**Descrição**
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like alignment tool
fRNA	functional RNA
GO	Genes Onthology
HMM	Hidden Markov Model
HTML	Hypertext Markup Language
IDRM	Intradermorreação de Montenegro
IMM	Interpolação de Modelos de Markov
LC	Leishmaniose Cutânea
LTA	Leishmaniose Muco-cutânea
LTA	Leishmaniose Tegumentar Americana
miRNA	micro RNA
ncRNA	Non coding RNA
NRDR	The Non-coding RNA Database Resource
ORF	Open reading frame
RNA	ribonucleic acid
rRNA	ribossomal RNA
SGBD	Sistema de Gerenciamento de Banco de dados
siRNA	small interfering RNA
snmRNA	small nom-mRNA
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
stRNA	small temporal RNA
tRNA	transfer RNA

Capítulo 1

Introdução

"Quando você sonha alto, todos os passos parecem ser o primeiro."
– Projota

A leishmaniose é considerada pela Organização Mundial da Saúde (OMS) como uma das seis doenças infecciosas mais importantes, estando presente em 88 países. Esta antropozoonose atinge cerca de 12 milhões de pessoas. A sua incidência é de dois milhões de novos casos ao ano, com cerca de 350 milhões de pessoas sob o risco de adquirirem a leishmaniose [Ministério da Saúde Secretaria de Vigilância em Saúde 2007]. Os agentes etiológicos dessa patologia são protozoários do gênero da *Leishmania*. A principal espécie desse patógeno no Brasil é a *Leishmania braziliensis* [Wallis et al. 2012].

Esta espécie, mesmo sendo o principal agente etiológico de uma importante doença, não possui o seu genoma estudado em uma escala tão ampla como outros organismos da mesma família como o *Trypanosoma cruzi*. Estão disponibilizadas no GenBank [Benson et al. 2014] da *Leishmania braziliensis* cerca de: 7833 genes, 1491 *expressed sequence tag*(EST) e 1 genoma (Estatística realizada com dados disponíveis no GenBank em 07 de Novembro de 2015). O genoma publicado pelo Sanger Institute (MHOM / BR / 75 / M2904), foi utilizado pela primeira vez no trabalho de [Laurentino et al. 2004] e disponibilizado no Genbank. No entanto, nenhuma atualização foi feita desde então.

As espécies de *Leishmania* possuem a presença de genes policistrônicos e diferenças genômicas ocasionadas pela seleção natural. A *L. braziliensis* foi considerada a espécie com maior diferença genômica em comparação com as espécies: *L. major* e *L. infantum* [Peacock et al. 2007]. Um dos motivos dessa diferença é o fato de que alguns genes conservados sofreram modificações na sua localização genômica em relação as outras espécies e a *L. braziliensis*. Atualmente não existe no NCBI a

anotação e mapeamento de componentes não codificantes para essa espécie. Esses componentes são chamados de RNA's não codificantes. Eles possuem funções catalíticas e reguladoras em atividades nucleares nos organismos [Kelada et al. 2013]. É necessária a predição dessas regiões para a caracterização das funções dessas estruturas na *L. braziliensis*.

Para aumentar o conhecimento sobre o genoma da *L. braziliensis*, é necessário reanotar o genoma. Pois existem muitos genes hipotéticos e putativos que podem executar funções importantes no metabolismo. A reanotação permitirá a classificação dos genes com função desconhecida, utilizando metodologias distintas. A identificação dos genes hipotéticos irão auxiliar na compreensão da dinâmica da patologia, desenvolvimento de fármacos e vacinas, além de auxiliar em estudos de genes específicos.

1.1 Justificativa

O aumento de casos observados pelo Ministério da Saúde de Leishmaniose tegumentar americana (LTA), evidencia a importância de efetuar estudos sobre a estrutura e a dinâmica do patôgeno. Estão disponibilizadas atualmente no GenBank cerca de 30586 sequências de Leishmania braziliensis. O genoma publicado da *L. braziliensis* (MHOM / BR / 75 / M2904) apresenta 7833 genes dos quais 5232 genes hipotéticos, representando cerca de 66% dos genes (Estatística realizada com dados disponíveis no GenBank em 07 de Novembro de 2015). Esses genes hipotéticos não apresentam similaridade suficiente com outros genes do dataset utilizado para comparação. Desta forma, torna-se difícil definir a sua função. Este tipo de anotação por similaridade possui limitações, devido a necessidade de um *dataset* vasto e curado para garantir as informações obtidas, [Yandell e Ence 2012].

A reanotação permitirá a classificação dos genes com função desconhecida, utilizando datasets curados de proteína e a identificação de novos genes codificantes e não codificantes de proteínas utilizando preditores *ab initio*. Além disso, o procedimento de reanotação permitirá corrigir as anotações realizadas para os genes putativos ou tipo. A disponibilização das anotações geradas publicadas em um banco de dados relacional auxiliará estudos de resistência a droga e métodos de diagnóstico da doença. Será possível também identificar as funções dos RNA's não codificantes no organismo do patógeno, podendo assim melhorar o entendimento da disposição dos ncRNA no genoma e a interações entre proteínas.

1.2 Objetivos

1.2.1 Objetivo Geral

Realizar uma reanotação do genoma de L. braziliensis (MHOM / BR / 75 / M2904).

1.2.2 Objetivos Específicos

1. Predizer e anotar os genes codificantes de proteínas no genoma de *L.braziliensis.*
2. Predizer RNA's não codificantes no genoma da *L. braziliensis.*
3. Organizar os dados gerados em um modelo de banco de dados relacionais.
4. Desenvolver uma *interface web* que possibilite a consulta aos dados gerados.

1.3 Organização do trabalho

Este trabalho esta organizado em sete seções: Introdução, Revisão Bibliográfica, Metodologia, Resultados, Discussões, Considerações Finais e Referências. Na Introdução o leitor entra em contato com o problema de estudo deste trabalho e é apresentado uma breve contextualização na temática que será mais profundamente abordada na próxima seção. Foi feita uma revisão bibliográfica sobre os principais temas que limiam este trabalho e se tormam necessários para o entendimento do mesmo.

Na Metodologia, descrevemos o processo utilizado para alcançar o objetivo esperado. Na seção de Resultados, estão descritos todos os resultados obtidos com a aplicação dessa metodologia. Na seção discussões é onde ocorre a discussão dos resultados. Na seção de considerações finais, são conclusões e informações analíticas sobre os resultados do projeto. Na seção Referências são encontradas os trabalhos e fontes que embasaram esta pesquisa. Como sugestão a leitura deve ser realizada na seguinte ordem: Introdução, Justificativa, Resultados Parciais, Fundamentação Teórica, Metodologia, Discussões e Considerações Finais.

Capítulo 2

Revisão Bibliográfica

"Pra quem tem pensamento forte o
impossível é só questão de opinião."
– Chorão E Thiago Castanho

2.1 *Leishmania braziliensis*

A leishmaniose é uma patologia classificada por seus aspectos clínicos em visceral ou tegumentar [Penna 1934]. A Leishmaniose Tegumentar Americana (LTA) apresenta duas manifestações clínicas clássicas: leishmaniose cutânea e leishmaniose mucosa ou muco-cutânea [Moreira 1895]. A leishmaniose cutânea (LC) se apresenta das seguintes formas: cutânea localizada, cutânea disseminada, recidiva cútis e cutânea difusa, enquanto que a forma muco-cutânea (LMC): mucosa tardia, mucosa de origem indeterminada, mucosa concomitante, mucosa contígua e mucosa primária.

Os vetores dessa patologia são insetos flebotomíneos pertencentes a família *Psychodidae*, Subfamília *Phlebotominae*, Gênero *Lutzomyia*. No Brazil, seis espécies de mosquito, são responsáveis pela transmissão da leishmaniose: *Lutzomyia flaviscutellata*, *L. whitmani*, *L. umbratilis*, *L. intermedia*, *L. wellcome* e *L. migonei*, [Ministério da Saúde Secretaria de Vigilância em Saúde 2007]. A espécie do mosquito está diretamente ligada ao local da transmissão do patógeno. Por exemplo as espécies *L. fisheri* e *L. neivai* são encontradas normalmente em ambientes domiciliares em áreas de transmissão da doença.

No ciclo de transmissão desse patógeno, existe uma interação reservatório-parasito, que é bastante complexa. Os reservatórios são animais que promovem a circulação do parasito na natureza. Infecções por *Leishmania* já foram descritas em diversas espécies de animais domésticos e silvestres como por exemplo os cães, gatos, camundongos e marsupiais [Gontijo e Melo 2004]. Normalmente, o ciclo de transmissão

ocorre com a infecção do vetor por meio do hospedeiro e a inoculação do parasito em seres humanos ou outros animais.

O diagnóstico da leishmaniose é realizado de forma direta com a observação do parasito ou da forma indireta através da intradermorreação de Montenegro (IDRM), também chamado de "Teste de Montenegro". O IDRM é uma reação de sensibilidade tardia, executado com a inoculação de antígenos da leishmaniose na face interna do antebraço. Após 48 a 72 horas é realizada a leitura da reação, para verificar o diâmetro das endurações. Caso possuam tamanho superior a 5mm são considerados positivos [Maria et al. 2010].

Esta patologia tem como agente etiológico os protozoários flagelados da família *Trypanosomatidae* do gênero *Leishmania*. As leishmanias são parasitos obrigatórios causadores da leishmaniose. No Brasil já foram identificadas sete espécies de *Leishmania*, que são: *L. (V.) braziliensis*, *L.(V.) guyanensis*, *L.(L.) amazonenses*, *L. (V.) lainsoni*, *L. (V.) naiffi*, *L. (V.) lindenberg* e *L. (V.) shawi*. Segundo o estudo realizado por [Wallis et al. 2012], utilizando uma amostra de 55 pacientes de diversas regiões do Brasil, foi avaliado a predominância de espécies de *Leishmania* em lesões cutâneas da leishmaniose no território nacional. Como resultado desse estudo foi demonstrado que a *Leishmania braziliensis* é a principal espécie responsável pelos casos de Leishmaniose tegumentar americana no Brasil. Na Figura 2.1, a distribuição geográfica das espécies de *Leishmania* encontradas no Brasil demonstra a presença da *L. braziliensis* em todo o território nacional.

Figura 2.1: Distribuição das espécies de *Leishmania* causadoras da leishmaniose tegumentar americana.Fonte: [Ministério da Saúde Secretaria de Vigilância em Saúde 2007]

A *L. braziliensis*, tem o seu ciclo vital alternado entre o estado de forma não flagelada, também chamado de amastigota e flagelada, também chamado de promastigota [Marsden e Nonata 1975]. A forma amastigota, normalmente é localizada em tecidos dos hospedeiros vertebrados. A promastigota é localizada normalmente no estômago do transmissor [Ministério da Saúde Secretaria de Vigilância em Saúde 2007]. A *Leishmania braziliensis* é uma espécie do sub-gênero *Viannia*. Este subgênero tem como característica principal espécies que ocasionam a Leishmaniose Tegumentar. Em 2004, o genoma da *Leishmania braziliensis* foi sequenciado no trabalho de [Laurentino et al. 2004]. Este trabalho utilizou uma metodologia de sequenciamento de nova geração, identificando que o genoma desta espécie é organizado em 35 cromossomos, totalizando cerca de 10.848 *reads* que foram analizados utilizando o *software* PHRED/PHRAP/CONSED [Ewing e Green 1998]. No entando, desde então nenhuma atualização foi realizada na anotação deste genoma.

2.2 Predição de elementos genéticos

2.2.1 Predição Gênica

A predição gênica é o processo de analisar o genoma, afim de identificar possíveis regiões gênicas [Yandell e Ence 2012]. Na década de 1990, ocorreram avanços neste processo, utilizados até hoje em técnicas mais atuais [Guigó et al. 1992][Solovyev et al. 1994]. Neste período surgiu uma abordagem para identificação de genes, fácil e rápida conhecida como *ab initio* [Zhu et al. 2010]. Existem diversas ferramentas com variados modelos matemáticos afim de melhorar a acurácia do processo. Algumas delas são: Glimmer [Kelley et al. 2012], Easy Gene [Larsen e Krogh 2003] e GeneMarkS [Besemer et al. 2001].

Um preditor gênico utilizado conhecido pela sua alta sensibilidade na detecção de genes em diversas espécies é o GLIMMER, [Delcher et al. 2007]. Esse preditor alcançou 99% de sensibilidade para genomas de bactérias, archaea e vírus e outros eucariotos com alta densidade genômica. A versão do 3.0, é a mais recente demonstrando melhora na acurácia em relação as anteriores. A melhora é devido a alterações no modelo de interpolação de Markov do algoritmo, [Majoros et al. 2004].

Para predição gênica esse algoritmo gera todas as *substrings* possíveis do dataset. Esses padrões são tabulados em seis *frames* de leitura. Para simplificar o cálculo das frequências é analisado um único *frame* de leitura por vez para denotar o número de ocorrências de uma *string* (sequência), $S = s1, s2, s3, ..., sn$. Esse procedimento é repetido para cada um dos seis *frames* de leitura e são calculados as estimativas iniciais de probabilidade da base Sx ocorrer no contexto Sx, i, (x, representa a base e o i, representa a posição). A probabilidade da base é calculada pelo algoritmo utilizando a fórmula representada na Figura 2.2.

$$P_i(S_x) = P(S_x|S_{x,i}) = \frac{f(S_{x,i})}{\sum_{b\epsilon\{atgc\}} f(S_{x,i}, b)}$$

Figura 2.2: Fórmula matemática que calcula a possibilidade da base S_x ocorrer na posição i. Fonte: [Salzberg et al. 1998]

Existe uma medida que é associada ao $P_i(S_x)$ para mensurar a confiabilidade da acurácia de um determinado valor ser verdade, esta medida é representada por $\lambda_i(S_x)$. Dois fatores influenciam essa medida, o primeiro é a simples frequência de ocorrência no contexto da string (S_x). Se esse valor assumir um número superior ao limite configurado para dados de treinamento, o valor do $\lambda_i(S_x)$ é modificado para 1.0. Para calcular a probabilidade da próxima base o algoritmo leva em consideração as frequências relativas (FR) na *string* de contexto (S_x).

A FR é calculada para cada base de nucleotídeos formando os valores das probabilidades da Interpolação de Modelos de Markov (IMM). Desta forma verifica-se por meio do maior valor do λ_i, adotando-a então como a melhor predição para a próxima base. É calculado a confidência do teste $X^2(c)$, nos casos em que as frequências não consistirem com as probabilidades do IMM. Pode ser visualizado a fórmula completa do cálculo de $\lambda_i(S_x)$ na figura 2.3. O valor λ_i, assume os valores das probabilidades $IMM_i(S_{x,i,b})$, sendo $b\epsilon\{atgc\}$.

$$\lambda_i(S_{x-1}) = \left\{ \frac{c}{400} \sum b\epsilon\{acgt\} f(s_1, s_2 ... s_i b) \right.$$

Figura 2.3: Fórmula para calcular a confiabilidade das predições feitas utilizando o IMM. Fonte: [Salzberg et al. 1998]

O GLIMMER necessita de evidências externas para a identificação de genes ou determinação de estruturas de íntron e éxons, [Brent 2005]. Essas evidências são padrões para a identificação dessas estruturas. Muitos preditores dessa metodologia já possuem seus parâmetros pré-calculados para alguns genomas clássicos como por exemplo o *Homo sapiens* e a *Arabidopsis thaliana*. Porém quando não se tem o preditor treinado para um organismo específico, o preditor executa o processo de treinamento para este novo genoma. Existem *pipelines* que simplificam o treinamento de preditores, como por exemplo, o *pipeline* MAKER, que treina o preditor Augustus [Stanke e Waack 2003] e SNAP [Korf 2004].

Um outro preditor gênico utilizado em projetos de anotação genômica é o GENSCAN [Burge e Karlin 1997]. Este algoritmo tem por objetivo a identificação de genes em organismos eucariotos e unidades de éxons, íntrons, regiões de splice e promotoras.

Este algoritmo utiliza a técnica de *Hidden Markov Model* (HMM) para efetuar suas predições. Essa técnica supõe que os estados são observados apenas indiretamente.

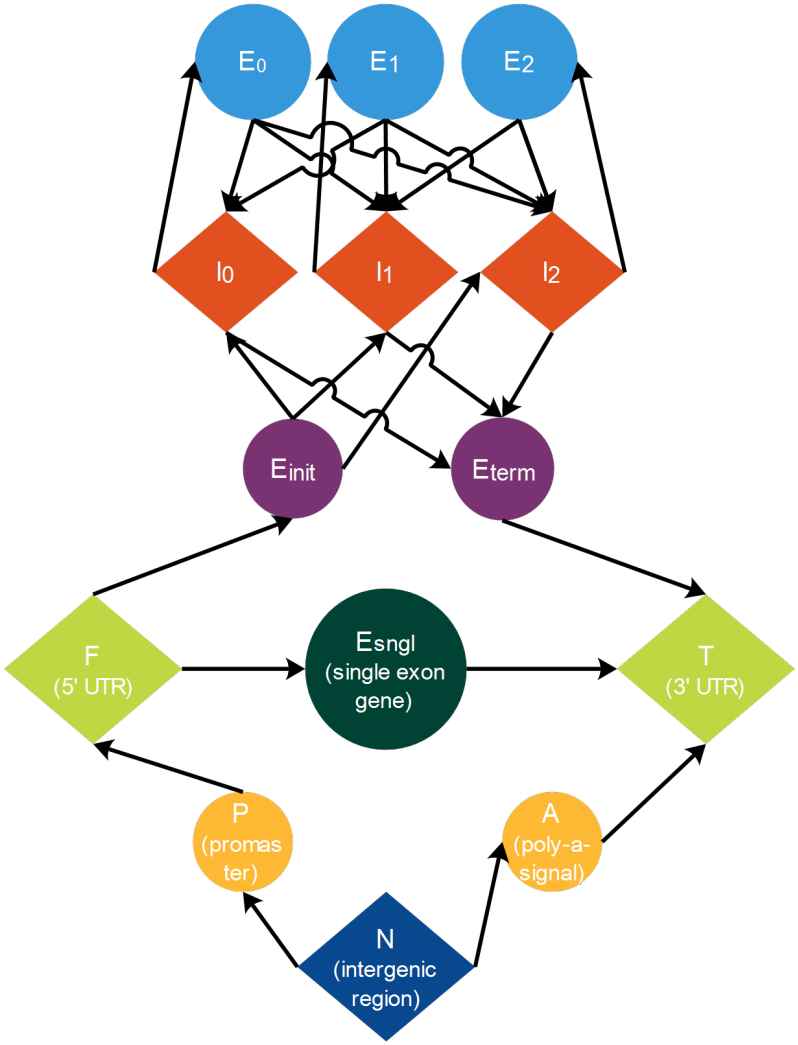

Figura 2.4: Adaptação do modelo geral de estados das estruturas genômicas. Fonte: [Burge e Karlin 1997]

Estes estados são representações de estruturas genômica de eucariotos: éxons, íntrons, regiões intergênica e etc. Inicialmente o algoritmo utiliza o modelo geral de estados das estruturas genômicas, que pode ser visualizado na figura 2.4. Baseado no *data set* entregue como parâmetro, calcula-se as probabilidades individuais para estado de transição do modelo da figura 2.5.

A equação da figura 2.5, descreve o cálculo de probabilidade da sequência estar em um determinado estado do modelo da figura 2.4. O avanço trazido pelo utilização dos HMM's é a avaliação de mais estados como: regiões intergênicas, promotores, UTRs, éxons e íntrons. Porém, apenas os genes codificantes, não é suficiente para entender a dinâmica do patógeno. É necessário mapear e predizer também os RNA

$$P(\varphi|S) = P(s_1|q_1)T(q_1|q_2)P(s_2|q_2)...T(q_{n-1}|q_n)P(s_n|q_n)P_0(q_n)$$

Figura 2.5: Fórmula da probabilidade de uma determinada sequência estar no estado mais provável (φ). Fonte: [Zhang 2002]

não codificantes . Em trabalhos anteriores já foi demonstrado que eles poderiam assumir funções nucleares altamente específicas e interações em catalizações complexas, [Kelada et al. 2013].

2.2.2 Predição de RNAs não codificantes

A aplicação de novas técnicas de avaliação da expressão (micro arranjo), tem detectado muitos grupos de RNAs não codificantes. Define-se como RNAs não codificantes (ncRNA ou RNAnc), o RNAs que não é codificado em proteína, [Dias Correia e a.a Dias Correia 2007]. Porém, sabe-se muito pouco sobre a função desses ncRNA. Existe alguns trabalhos demonstrando que algumas dessas estruturas possuem funções estruturais e regulatórias no processo de expressão de mRNA's, [Kelada et al. 2013]. Os ncRNA são classificados na literatura em: fRNA (*functional RNA*), miRNA (micro RNA), rRNA (ribossomal RNA), siRNA (*small interfering RNA*), snRNA (*small nuclear RNA*), snmRNA (*small nom-mRNA*), snoRNA (*small nucleolar RNA*), stRNA (*small temporal RNA*) e tRNA (*transfer RNA*), [Paschoal et al. 2012]. Existem trabalhos que assossiam essa estrutura ao spectrum de doenças. O HMDD (`http://www.cuilab.cn/hmdd`), armazena cerca de 400 spectrum de doenças (doenças humanas incluindo câncer e doenças cardiovasculares) que foram reportadas com associações a miRNA (microRNA), [Li et al. 2014].

A predição dessas estruturas é possível utilizando preditores modernos. É comum em processos desse gênero, se utilizar um conjunto de *softwares* para efetuar a predição dos ncRNA. Um desses *pipelines* é o structRNAfinder (http://integrative bioinformatics.me/structrnafinder/). Utilizando a metodologia de novo, este *pipeline* foi desenvolvido em perl, javascript e html.

O structRNAfinder inicia o seu processo construindo os modelos de covariância utilizando os *softwares* cmbuild e cmcalibrate, [Nawrocki e Eddy 2013]. Um modelo de covariância é o conjunto de modelos estatísticos que descrevem estruturas secundárias e consensus de estruturas primárias de RNA, [Eddy e Durbin 1994]. Após a obtenção do modelo, utiliza-se os programas, cmscan e cmsearch disponibilizados no Infernal [Nawrocki e Eddy 2013]. Para efetuar a busca de um modelo de co-variância contra um banco de dados de sequência e o inverso também.

Após a busca são filtrados os melhores hits para a extração de uma sequência madura. Com o dataset de sequências maduras utiliza-se o algoritmo RNAfold, para converter e colorir as estruturas de RNA da sequência *query*. Nesta etapa, também são geradas

as fig das estruturas de RNA para o relatório final. Após esta etapa são utilizados os datasets do RFAM [Burge et al. 2013], para buscar indícios taxonômicos nas sequências e por fim gera-se um relatório em HTML (*HyperText Markup Language*).

2.2.3 Algoritmos de busca por similaridade

Durante o processo de anotação das sequências de um determinado organismo, em muitos momentos se faz necessária a comparação entre sequências. Esses algoritmos também são chamados de algoritmos de alinhamento. Pode-se classificar esses alinhamentos em globais e locais. Os alinhamentos locais são utilizados para buscar apenas as regiões de alta similaridade, não importando as regiões adjacentes. Já os alinhamentos globais, buscam o alinhamento completo das sequências analizando-as por completo. Normalmente, os processos de anotação de sequência utilizam algoritmos de alinhamento local, como por exemplo o BLAST [Altschul et al. 1990], o BLAT, [Kent 2002] e o Bowtie [Langmead et al. 2009].

2.2.4 BLAST

O BLAST (*Basic Local Alignment Search Tool*), é um algoritmo que baseia-se em um modelo heurístico para efetuar o alinhamento da sequência com um banco de dados de sequências, [Altschul et al. 1990]. Este algoritmo é utilizado na realização de buscas por similaridade em sequências nucleotídicas e proteínas devido: ao baixo custo computacional, escalonabilidade entre as configurações de hardware e sua consolidação na área. Atualmente o BLAST foi especificado e foram desenvolvidos variações, para cada tipo de busca. São eles: blastn, blastp, blastx, tblastn e o tblastx, [Johnson et al. 2008].

O blastn é um *software* utilizado para efetuar buscas em bancos de dados de nucleotídeos utilizando uma sequência também de nucleotídeos. O blastp é a variação do BLAST utilizada para realizar buscas em bancos de dados de proteínas utilizando uma sequência de proteína. O blastx é utilizado para a comparação de banco de dados de proteínas utilizando uma sequência de nucleotídeo traduzidos. O tblastn é a variação do BLAST que efetua buscas entre bases de dados de sequências de nucleotídeos traduzidos e uma sequência de proteína de referência. O tblastx é utilizado para efetuar buscas por similaridade entre uma base de nucleotídeos traduzidos e uma sequência de nucleotídeos traduzidos como referência. As variações supracitadas são utilizadas baseadas no tipo de banco de dados e sequência de referência que serão utilizados na busca por similaridade.

No processo de alinhamento este algoritmo utiliza como parâmetro duas sequências ou um banco de dados de sequências. Uma dessas sequências é chamada de *query*, a outra sequência ou um banco de dados de sequência é utilizado para busca de similaridade, chamado de subject. Para início do processo, o algoritmo busca uma

subsequência de tamanho ou tamanho da palavra (W) na região central da sequência de referência. Após a definição da subsequência, o dataset é dividido em um *subdataset* com todas as combinações possíveis de sequências com tamanho igual a W, [Altschul et al. 1990].

São calculados então a soma da pontuação das regiões candidatas a similaridade utilizando como base a matriz de *score*. Após o cálculo das pontuações, as sequências que obtiveram valor inferior ao ponto de corte (T) saem do processo de expansão do alinhamento. Então o W aumenta cerca de 1 W para cada lado da subsequência inicial e o processo é todo refeito, até não existir mais sequências com uma pontuação superiar ao T. O tamanho inicial do W no blastn é de 11, [Altschul et al. 1990]. O BLAST compara todas as sequências com o dataset e calcula o *E-value*, utilizando a seguinte fórmula:

$$E = mnp$$

O m representa o total de número de resíduos do dataset. O n representa o número total de resíduos da sequência query e o p representa a probabilidade do alinhamento ser randômico. Muitos autores apontam como vantagens desse algoritmo o seu baixo custo computacional e a velocidade do alinhamento executado com ele. Surgiram com a avanço das ferramentas tecnológicas alguns algoritmos que fizeram melhorias no BLAST, como por exemplo o *BLAST Alignment Tool* (BLAT).

2.2.5 BLAT

O BLAT foi desenvolvido para trabalhar com sequências de mRNA (RNA mensageiro) e DNA, [Kent 2002]. Este algoritmo tem as suas etapas similares as do BLAST, no entando existem algumas otimizações para aumento de velocidade e redução de tempo nas buscas e alinhamentos executados. Assim como o BLAST, o BLAT pode ter o seu procedimento dividido em duas etapas: busca e alinhamento. Na etapa de busca o algoritmo encontra regiões entre duas sequências que possam ser homologas. A segunda etapa é chamada de alinhamento, é nela que o algoritmo monta os alinhamentos entre as regiões afim de definir por critérios quais as regiões são realmente homólogas.

Inicialmente o BLAT executa a etapa de busca por possíveis regiões homologas. Este procedimento inicialmente efetua a procura por sequências de tamanho k no dataset e comparando com a sub-sequência da janela de tamanho w, sendo w um valor definido pelo *software* ou pelo usuário. Algumas implementações dessa busca possuem um cálculo de probabilidade para verificar a chance de uma sub-sequência ser similar. As sequências que não atingem um determinado percentual (P1) é retirada do dicionário de possíveis homologias. O cálculo de P1é feito com base na seguinte formula (M representa a taxa de similaridade entre as áreas homólogas e o k representa o tamanho da subsequência):

$$P1 = MK$$

Está medida reduz o custo computacional da próxima etapa, tendo em vista que sequências com baixa similaridade não passarão pelo processo de alinhamento. Após a montagem do dicionario de regiões com possíveis homologias, ocorre então a etapa de alinhamento. Nesta etapa o BLAT utiliza um algoritmo de alinhamento chamado de Intronerator (`http://www.csc.ucsc.edu/~kent/intronerator`) [Kent e Zahler 2000]. Este algoritmo inicia gerando uma lista de hits entre a sequência de referência e o banco de dados de regiões homólogas. Após gerar essa lista de hits de tamanho w e alinhá-lo os alinhamentos são expandidos até encontrar regiões não homólogas. Após está etapa é gerado um arquivo de texto com os dados gerados no processo para posterior analise. Todos os dados gerados e os que são utilizados como entrada para esses algoritmos, precisam ser armazenados para que possam ser recuperados depois e servir como fundamento e insumos para outras pesquisas.

2.3 Anotação Gênica

Anotação genômica é um termo utilizado para definir o processo exploratório do genoma afim de identificar a localização e a função de genes e outras estruturas genômicas. Classifica-se anotação genômica em: estrutural e a funcional. A anotação genômica estrutural é conhecida como o processo de identificação dos genes, íntrons e éxons. Diferentemente a anotação genômica funcional é definida como o processo de associação das informações funcionais de ontologia com anotações estruturais, [Yandell e Ence 2012].

O processo de anotação é atualmente é dividido em duas fases: A fase computacional e a fase de consolidação e sintetização dos dados. Na fase computacional são aplicados diversos programas em conjunto para fazer a identificação de estruturas gênicas e a função dessas estruturas no organismo foco do estudo. A fase de consolidação e sintetização dos dados é a etapa em que os dados gerados pelos *softwares* são analisados e estruturados em um repositório padronizado.

A rotina de utilização do conjunto de *softwares* para obter um determinado resultado é conhecida como *pipeline*. São utilizados muitos algoritmos durante a anotação. O processo de condensação dos dados gerados em um banco de dados estruturado, ainda é um desafio. A dificuldade desse processo se deve a não padronização das saídas dos programas e a variabilidade dos *pipelines*, tornando árduo o processo de consolidação.

2.3.1 Anotação Gênica na atualidade

Atualmente a anotação genômica é realizada de duas formas: via plataforma de anotação ou via pipeline (conjunto de algoritmos). A plataforma de anotação é basicamente uma ferramenta facilitadora do processo de anotação. Nessas ferramentas normalmente existe uma interface que facilita a aplicação dos algoritmos de predição e os alinhamentos para identificação de funções. Muitas vezes essa ferramenta utiliza um pipeline de algoritmos. Uma plataforma de anotação genômica é a Sabiá, [Almeida et al. 2004].

Esta ferramenta permite a anotação integrada de genomas bacterianos. A anotação no Sabiá é realizada com a aplicação de diversos algoritmos de predição (GLIMMER) e a realização de alinhamentos entre os genes preditos e bases de dados como o KEGG ou SWISSPROT. Após estas análises é realizada a anotação manual e revisão do gene. Essa plataforma desenvolvida pelo LNCC (Laboratório Nacional de Computação Científica) é *open-source*, porém é necessária a solicitação do código.

Uma outra metodologia é a utilização de *pipelines* de anotação. *Pipelines* são algoritmos interligados que executam todo o processo de anotação de uma forma computacionalmente automática. Um exemplo de *pipeline* de anotação genômica é o MAKER, [Cantarel et al. 2008]. Esta ferramenta é compatível com o padrão de armazenamento GMOD (*Generic Model Organism Database*). Esse processo foi utilizado na anotação do banco de dados Flybase [McQuilton et al. 2012].

2.4 Armazenamento de dados biológicos

Um banco de dados é uma coleção de dados que possuem no minimo uma relação semântica, [Elmasri e Navathe 2011]. Entende-se por dados fatos que podem ser registrados. No caso de dados biológicos podem ser: sequências de DNA, proteínas, biomarcadores, resultados de análises, dados clínicos ou dados epidemiológicos. Atualmente, os bancos de dados biológicos são classificados pelo tipo de dado que eles armazenam, que pode ser dos tipos: primários, secundários e terciários ou especializados.

Os bancos primários são repositórios de dados biológicos brutos, que são normalmente sequências de DNA / RNA, proteínas, dados clínicos em geral. Alguns exemplos de repositórios primários são: GenBank Nucleotide [Benson et al. 2014] e o NRDR [Paschoal et al. 2012]. Este tipo de banco de dados é utilizado normalmente como parâmetro para análises. Os bancos secundários, são repositórios de dados biológicos secundários ou que possuem dados resultantes de um processamento analítico ou modificador. Alguns exemplos dessa categoria de bancos biológicos são: HIV Database (`http://www.hiv.lanl.gov`) e o FlyBase [McQuilton et al. 2012].

Os bancos de dados biológicos especializados, são descritos por alguns autores como um banco de dados secundário, específico para um determinado organismo, es-

pécie ou categoria. Um exemplos desse tipo de banco de dados é o TrytripDB (`http://tritrypdb.org/tritrypdb/`). Que é um banco especializado em patógenos da família *Trypanosomatidae*. Os bancos de dados biológicos, mesmo com a classificação enquanto armazenamento de dados, podem ter a metodologia computacional variada. Uma dessas metodologias computacionais para o armazenamento de dados é o modelo relacional de dados.

2.4.1 Modelos relacionais de dados

O modelo relacional de dados, diferentemente do *Flat file* (armazenamento em uma única tabela), organiza os dados em múltiplas tabelas. A ligação semântica entre duas ou mais entidades (tabelas), é chamada de relação. Uma tabela no modelo relacional de dados é chamada de entidade, pois é normalmente uma representação de uma entidade real do programa para o qual aquele banco foi modelado. Esta tabela é formada por colunas que representam os atributos ou características da entidade e as linhas que representam instâncias de armazenamento de dados daquela determinada entidade. Pode-se observar na Figura 2.6, um exemplo de banco de dados modelado de forma relacional e os mesmos dados armazenados em um banco de dados *Flat file*.

Flat File

Name, States, Course number, Course name|John Smith, Texas, Biol 689, Bioinformatics|Jane Doe, Kansas, Bich 441, Biochemistry|William Brown, Illinois, Chem 289, Organic Chemistry|Jennifer Taylor, New York, Hort 201, Horticulture|Howard Douglas, Texas, Math 172, Calculus

Table A

Student #	Name	State
1	John Smith	Texas
2	Jane Doe	Kansas
3	William Brown	Illinois
4	Jennifer Taylor	New York
5	Howard Douglas	Texas

Table B

Student #	Course #
1	Biol 689
2	Bich 441
3	Chem 289
4	Hort 201
5	Math 172

Table C

Course #	Course name
Biol 689	Bioinformatics
Bich 441	Biochemistry
Chem 289	Organic chemistry
Hort 201	Horticulture
Math 172	Calculus

Figura 2.6: Demonstração gráfica do armazenamento de dados no formato Flat File e em um modelo relacional. Fonte: [Xiong 2006]

Devido a estruturação dos dados, este modelo é capaz de executar consultas mais complexas do que o modelo *Flat file*. A complexidade e tamanho de um banco de dados biológico seguindo este modelo varia conforme a estrutura biológica ou dados que serão armazenados. Para fazer consultas em uma base de dados relacional, utiliza-se uma linguagem desenvolvida para essa finalidade, conhecida como SQL (*structured query language*). Esse tipo de modelo de dados necessita de um sistema de gerenciamento de banco de dados, que é um conjunto de programas que permitem a administração e manutenção das bases, [Elmasri e Navathe 2011].

Existem algumas vantagens de se utilizar uma abordagem de armazenamento de dados em Sistema de Gerenciamento de Banco de dados (SGBD). Devido a utilização de técnicas mais modernas de manutenção e acesso aos dados, bancos de dados relacionais tendem a ter um menor custo computacional do que os *Flat files*. Essa melhoria ocorre principalmente a técnicas utilizadas para melhorar o acesso aos dados armazenados: índices e outras técnicas como manter os dados maiores em memória para otimização de consultas. Um problema comum para outros tipos de armazenamento é a redundância de dados. Em Flat files é comum ocorrer o armazenamento redundante, que é o mesmo dado ser armazenado várias vezes em lugares diferentes.

Devido ao sub-sistema do SGBD de controle de redundância e a necessidade de projetar o banco de dados antes de criá-lo, este problema é raro ocorrer nesse tipo de modelo. Principalmente se modelo do banco de dados for normalizado, garantindo assim a não ocorrência de redundância ou caso necessário um controle das redundâncias necessárias. Uma outra vantagem é a segurança de acesso aos dados armazenados. O SGBD garante um controle de acesso feito através de usuário e senha para proteger os dados e garantir que apenas pessoas autorizadas terão acesso. Este controle não existe no *Flat file*, permitindo que qualquer usuário do computador, abra e edite os dados armazenados.

O uso de uma linguagem desenvolvida para pesquisa nesse tipo de repositório facilitou o desenvolvimento de um subsistema do SGBD para a execução e otimização das consultas. Com essa tecnologia é possível executar consultas mais completas sem o problema de concorrência de acesso ao arquivo. O problema de concorrência é comum em sistemas multiusuários baseados em *Flat file*. Normalmente os SGBD's possuem um sistema de *buffering* ou *caching* que mantém na memória principal, partes dos dados armazenados, afim de agilizar e melhorar a performance das consultas.

2.5 Genome Browser

Após todo o processo de anotação genômica da *Leishmania braziliensis* e armazenado os dados no banco de dados gerados, necessita-se ter uma ferramenta que permita a visualização dos dados. Esse tipo de ferramenta nos permite uma visualização de anotação genômica e estruturas mapeadas. Existem diversos programas com essa objetivo, alguns deles são: Artemis, UCSC e JBrowse.

2.5.1 Artemis

O Artemis é um *software* de visualização de informações da biologia molecular moderna, [Rutherford et al. 2000]. Ele foi escrito em Java e tem por característica

principal ser um programa *open source*. Como requisito mínimo para a execução desse algoritmo é apenas um computador com JVM (Java Virtual Machine).

Este *software* teve o seu desenvolvimento preoculpado com a interoperabilidade permitindo a execução do Artemis em plataformas utilizando UNIX, Mac OS X ou Windows. Esta facilidade na execução do algoritmo, tornou o Artemis um visualizador muito utilizado em diversos projetos de anotação genômica. Atualmente o Artemis permite a integração da anotação com o banco de dados EMBL. Esse cruzamento das informações permite utilizar dados já conhecidos e integrá-los a uma nova anotação do organismo.

Na sua versão mais recente foi adicionado no Artemis a função de visualização e análise de dados provenientes das novas tecnologias de sequenciamento,[Carver et al. 2012]. O funcionamento desse sistema é bem simples para o usuário. Para visualizar os dados de anotação é necessário apenas do arquivo (.gbk). Porém é aceitado também arquivos de alinhamentos como por exemplo os arquivos (.bam) e (.sam). A identificação das estruturas e a montagem dos mapeamentos no genoma é feito automaticamente após o arquivo ter sido adicionado. Existe uma interface implementada para permitir a plotagem de dados armazenados em bancos de dados relacionais que sigam o padrão do Artemis.

Na figura 2.7, é observada a simplicidade da interface gráfica do Artemis. A interface de fácil uso é um ponto positivo no Artemis, porém o fato de não ser um sistema *web* é uma desvantagem desse *software*.

2.5.2 UCSC Genome Browse

O *UCSC Genome Browse* é um programa *web* que permite a visualização de um genoma ou sequência com suas anotações mapeadas, [Karolchik et al. 2009]. Para gerar os gráficos, o *UCSC Genome Browse* utiliza scripts em JavaScript para montar o visualizador. Este *software* foi desenvolvido pela Universidade da Califórnia Santa Cruz (UCSC) inicialmente para facilitar a visualização das anotações dos datasets de genoma. Porém, com o tempo muitos projetos de anotação ao redor do mundo utilizaram o *software* para visualizar as anotações.

Basicamente o *UCSC Genome Browser* dispõe as anotações alinhadas em faixas ou *"tracks"*. Inicialmente era aceito os arquivos padrões de anotação (.SAM, .BED, .BAM e .GBK), porém na versão atual ele aceita bancos de dados MySQL. A integração com o MySQL é feita por uma ferramenta interna do programa chamada UCSC Table Browser. Para utilizar esse programa é necessário acomodar os dados em um modelo pré-definido pela documentação disponível no site da UCSC (http://genome.ucsc.edu). Este software permite a análise genômica de diversos aspectos e níveis, o que facilita o seu uso. Utilizando o recurso de zoom por exemplo é possível visualizar desde nucleotídeos a proteínas e anotações mapeadas em cada região. No site existe um exemplo da tela principal do genome browse que está representado na figura 2.8.

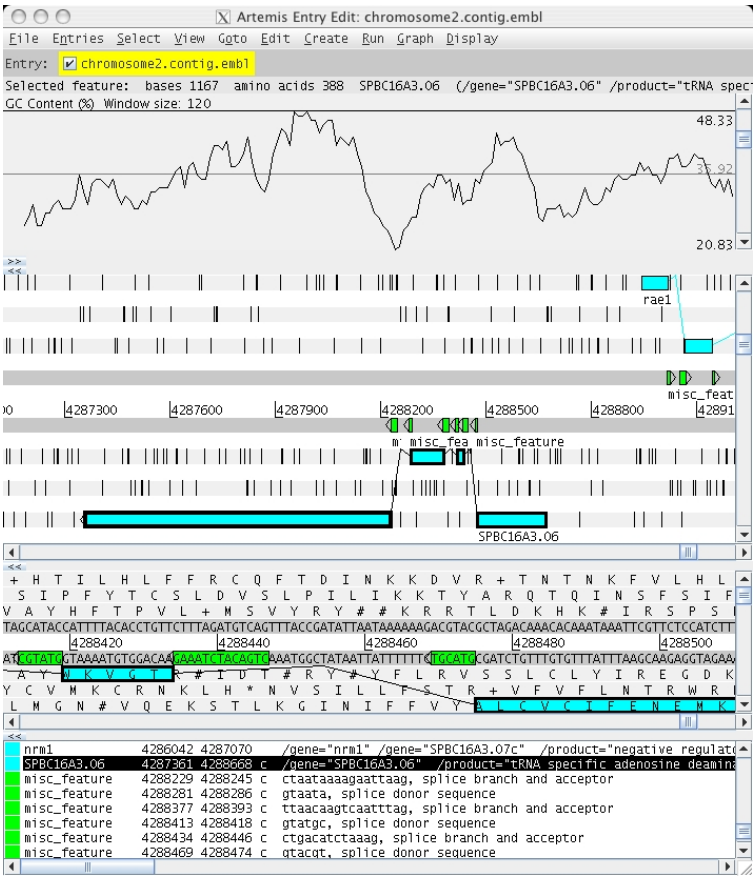

Figura 2.7: Interface do Artemis apresentada como exemplo no site. Fonte: https://www.sanger.ac.uk/resources/software/artemis/.

2.5.3 JBrowse

Um outro genome browse que surgiu após o *UCSC Genome Browse* foi o JBrowse, [Skinner et al. 2009]. O JBrowse procurou solucionar alguns problemas de performance do UCSC e possuir mais recursos implementados. Diferente dos anteriores apresentados, essa ferramenta utiliza JavaScript e HTML5 como base de programação. O código e manual do JBrowse estão disponíveis no site: `http://jbrowse.org/`. A interface dessa ferramenta implementa os padrões de responsibilidade, este fato faz com que ele se adapte a diversos formatos de telas. Este recurso não foi implementado nos demais citados nesse trabalho.

Apresentando uma interface simplificada e facilitada (demonstrada na figura 2.9), em comparação com o UCSC garantindo uma melhor performance. Esse teste de performance é apresentado no trabalho de [Skinner et al. 2009] comparando a quantidade de atualizações da página nas ações de zoom, reordenamento, adição e remoção de trilhas. O formato de arquivos padrão para o armazenamento de dados do JBrowse é o JSON. Porém existem scripts já desenvolvidos e disponíveis que convertem arquivos: .BAM, BIODB, SAM e GBK para o JSON. Estes scripts de conversão e

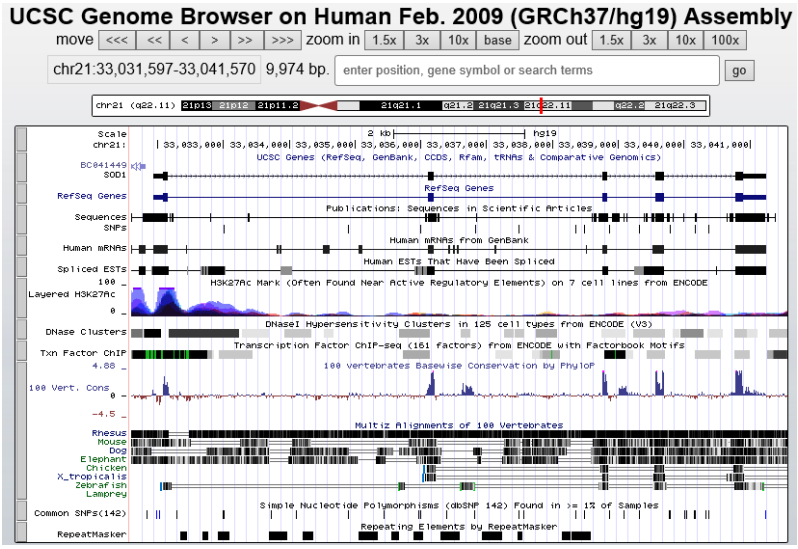

Figura 2.8: Interface do *UCSC Genome Browse* apresentada como exemplo no site. Fonte: http://genome.ucsc.edu

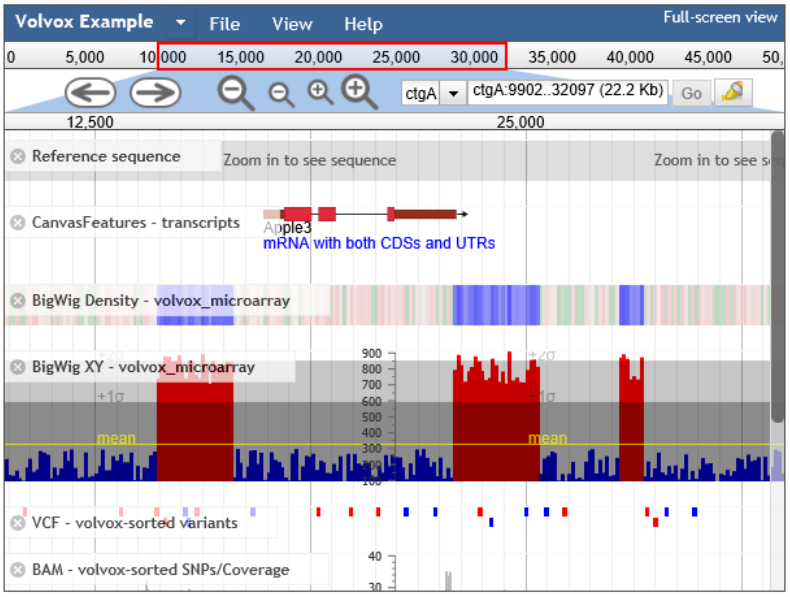

Figura 2.9: Interface do JBrowse apresentada como exemplo no site. Fonte: http://jbrowse.org/

todos os demais necessários para instalação estam disponíveis em PERL.

2.6 Métodos Estatísticos

Para analisar os dados gerados pela anotação, é necessário utilizar técnicas estatísticas. Existem muitas metodologias estatísticas para análise de dados, algumas delas são: Coeficiente de Correlação e Testes de Hipóteses.

2.6.1 Coeficiente de Correlação

Uma análise estatística que está presente em muitos trabalhos da literatura é a busca de correlação entre duas variáveis numéricas. Por exemplo é conhecido que o alto consumo de açúcar aumenta as chances do indivíduo desenvolver diabetes. Este exemplo mostra que existe uma relação entre as variáveis consumo de açúcar e probabilidade de desenvolver diabetes. Essa relação é mensurada estatisticamente através do coeficiente de correlação. Uma medida para esse coeficiente é o coeficiente de correlação de Pearson. A fórmula para o cálculo dessa medida é, [Vieira 2008]:

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\left[\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}\right]\left[\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{n}}\right]}$$

O coeficiente r indica o grau da correlação que varia no intervalo de $-1 \leq r \leq +1$. O valor de r igual a 1 indica uma correlação perfeita positiva e -1 indica uma correlação perfeita negativa. Para calcular o coeficiente de correlação em uma amostra é necessário que três requisitos sejam atendidos. O primeiro é que as unidades de medidas sejam uma amostra representativa da população. O segundo requisito é que cada unidade da amostra deve fornecer valores de X e Y, ou seja, toda unidade tem que ter mensurada a variável de cada um dos eixos. O terceiro e último é que as variáveis devem ser mensuradas de uma forma independente.

2.6.2 Testes Estatísticos

2.6.3 Teste T

Os testes estatísticos são também chamados de testes de hipóteses. Na prática eles testam hipóteses em uma determinada população. Normalmente são criadas duas hipóteses, a nula e a alternativa. A hipótese nula ou hipótese da nulidade, na maioria das vezes afirma a inexistência de diferença entre as amostras ou grupos. Por sua vez, a hipótese alternativa tenta contradizer a nula, [Vieira 2008]. O teste T é utilizado então para escolher entre as duas hipóteses. Para essa escolha é utilizado o p-valor (valor de probabilidade) para decidir qual das duas hipóteses será descartada. Existe uma convenção sobre a análise do p-valor:

$$p - valor < 0.05 = A hipotese nula deve ser rejeitada$$

No caso do p-valor ser menor que 0.05, diz-se que os resultados são estatisticamente significantes. O teste T pode ser aplicado em duas situações diferentes: quando os dados são pareados e quando as amostras são independentes. Entende-se por dados pareados, variáveis de uma mesma população amostral avaliadas em momentos diferentes. Porém para aplicar o teste T em dados pareados é necessário que além das mesmas variáveis, que ele tenha um sentido lógico. O algoritmo do teste T para testes não pareados é descrito por [Vieira 2008] como:

1. Deve-se estabelecer as hipóteses.
2. Escolha o nível de significância do teste.
3. Nos passos seguintes aplica-se alguns cálculos nas amostras:
   (a) Nesta etapa é realizado o cálculo das diferenças entre todas as observações pareadas:
   $$d = x_2 - x_1$$
   (b) Então, calcula-se a média dessas diferenças:
   $$\bar{d} = \frac{\sum d}{n}$$
   (c) Após o cálculo da média, o próximo passo calcula a variância dessas diferenças:
   $$s^2 = \frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}$$
   (d) Calcula-se então o valor de t. Este valor é associado a (n-1) graus de liberdade pela fórmula:
   $$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$$
   (e) Por fim é feita a comparação entre o valor absoluto do t calculado e o valor crítico. Quando o valor absoluto do t calculado for igual ou superior ao valor crítico, rejeite a hipótese de que as médias são iguais, respeitando o nível de significância estabelecido.

Esse algoritmo é utilizado quando as amostras são dados pareados. Caso não sejam pareados, o teste T as fórmulas alteradas mas mantém o mesmo fundamento. Dados não pareados ou também chamados de dados independentes, são dados que não possuem uma relação de temporalidade imbutida neles. Por exemplo a uma aplicação em que uma amostra contenha pacientes submetidos ao tratamento de HIV e a outra seja pacientes controles. O algoritmo para utilização do teste T na comparação de duas amostras independentes é descrito por [Vieira 2008] como:

1. É realizado o cálculo da média e variância de cada amostra.
2. Neste passo calcula-se a variância ponderada, utilizando a fórmula abaixo:

$$s_p^2 = \frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}$$

3. Calcula-se então o valor de t. Este valor é associado a $(n_1 + n_2 - 2)$ graus de liberdade pela fórmula:

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_p^2}}$$

4. Por fim é feita a comparação entre o valor absoluto do t calculado e o valor crítico. Quando o valor absoluto do t calculado for igual ou superior ao valor crítico, rejeite a hipótese de que as médias são iguais, respeitando o nível de significância estabelecido.

2.6.4 Teste de Mann-Whitney

Outro teste estatístico utilizado para a comparação de amostras de dados independentes é o Teste de Mann-Whitney. Este teste detecta significância estatística de diferença entre dois grupos de dados independentes. Ele é apontado como uma alternativa para o teste T. O algoritmo utilizado para apontar esse teste é:

1. Inicialmente adota-se que n_1 seja o tamanho de amostra do menor dos grupos comparados e n_2 o tamanho de amostra do maior dos grupos comparados;
2. Calcula-se então a estatística do teste utilizando a seguinte fórmula:

$$MW = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - T$$

n_1 e n_2 são os tamanhos das amostras e T é a soma dos postos do grupo menor.

Para interpretar o resultado do teste é feita a comparação entre MW e o percentil.

Capítulo 3

Metodologia

3.1 Predição de genes

O genoma utilizado nesse projeto foi o da Leishmania braziliensis (MHOM / BR / 75 / M2904), disponibilizado no banco de dados GeneDB (`http://www.genedb.org/`) e no NCBI (`http://www.genbank.com`) no formato FASTA (.fasta) e GENBANK (.gb). Demonstrado na tabela 1, o dataset do genoma da *L. braziliensis* é formado por 35 cromossomos.

Após o download de cada sequência prosseguimos para a predição genica utilizando o Genscan (utilizado para predição em eucariotos) e o GLIMMER (utilizado para predição em procariotos). Foram utilizados preditores de eucarioto e procarioto, para aumentar a sensibilidade em regiões policistrônicas. Para executar as predições com o GENSCAN, foi utilizado o servidor web publicado no seguinte endereço: `http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::genscan`. O algoritmo foi utilizado com parâmetro de organismo utilizando a opção HumanISO. Em seguida foi executado o GLIMMER 3.02, também pelo servidor web publicado que pode ser acessado pelo seguinte endereço: `http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi`. Foi utilizado no GLIMMER a topologia linear que é a indicada no website da ferramenta para esse tipo de organismo. Após a predição, os dados gerados foram importados para um modelo de banco em SQLite, representado na figura 8. Utilizando uma combinação de scripts em Python com SQL, foi gerado um novo arquivo com todas as sequências preditas como gene pelos preditores em formato FASTA.

Cromossomo	Acession ID	Tamanho (pares de base)
1	FR798975	234570pb
2	FR798976	294768pb
3	FR798977	367702pb
4	FR798978	450572pb
5	FR798979	458091pb
6	FR798980	520397pb
7	FR798981	572114pb
8	FR798982	392409pb
9	FR798983	564735pb
10	FR798984	572386pb
11	FR798985	566625pb
12	FR798986	462944pb
13	FR798987	624911pb
14	FR798988	657644pb
15	FR798989	615681pb
16	FR798990	698648pb
17	FR798991	704727pb
18	FR798992	685542pb
19	FR798993	744148pb
21	FR798996	697672pb
22	FR798997	651295pb
23	FR798998	808824pb
24	FR798999	841233pb
25	FR799000	722594pb
26	FR799001	1007251pb
27	FR799002	1174396pb
28	FR799003	1176127pb
29	FR799004	1192020pb
30	FR799005	1346262pb
31	FR799006	1572434pb
32	FR799007	1595807pb
33	FR799008	1525959pb
35	FR799009	2041438pb
36	FR799010	2725128pb

Tabela 3.1: Tabela com os dados das sequências que formam o dataset utilizado neste trabalho.Fonte: NCBI.

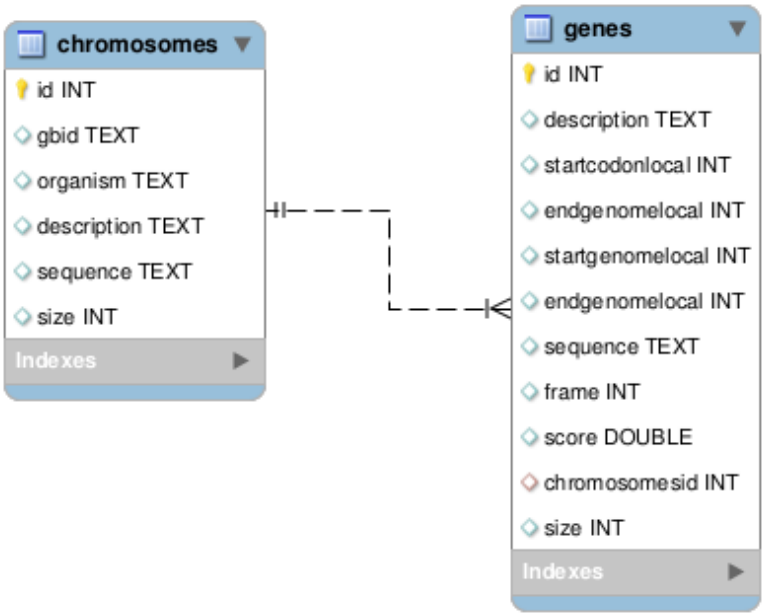

Figura 3.1: Representação gráfica do modelo relacional do banco de dados em SQLite utilizado para obter o arquivo FASTA com as predições gênicas. Fonte: Próprio Autor.

3.2 Identificação dos produtos e funções dos genes preditos

Com a unificação das predições em um arquivo FASTA, foi possível fazer a identificação e anotação dos genes preditos por meio de algoritmos de busca por similaridade em bases de dados de proteínas. Utilizou-se como banco de dados de referência de proteínas o SWISS-PROT (`http://www.uniprot.org/downloads/`), [Fern et al. 2015]. Para verificar a similaridade dos genes preditos com alguma proteína expressa, foi executado o BLASTx entre o banco de referência de proteínas e os genes preditos. Para executar o BLASTx é necessário formatar o SWISS-PROT para o formato de base de dados utilizado pelo BLASTx. Para esta tarefa utilizou-se o software do pacote NCBI-Blast+ chamado de MakeBLASTdb. Para execução, do BLASTx na base formatada foi utilizado o parâmetro (-evalue) para filtrar os alinhamentos que obtiveram valores do e-value superior ou igual a 1e-5, [Xiong 2006]. Os demais parâmetros não foram alterados, mantendo os valores padrões implementados no algoritmo BLASTx, [Altschul et al. 1990].

Foram consideradas como similares, hits com identidade acima de 50% no caso das proteínas e acima de 80% nas comparações de nucleotídeos, [Pearson 2014]. Obtendo o resultado do BLASTx, foi executada a identificação das proteínas encontradas por termo GO (*Gene Onthology*). Os termos GO são *index* de produto gênico padronizados para todas as espécies. A ontologia classifica as proteínas em três níveis: processo

biológico, componente celular e função molecular. Para esta tarefa de classificação foi utilizado a ferramenta web conhecida como AmiGO [Carbon et al. 2009]. Para classificar as proteínas identificadas o Amigo recebe como parâmetro todos os símbolos dos genes ou produto gênico das proteínas. Ele permite efetuar o filtro quais bancos de proteínas serão utilizados. Para esta análise limitamos ao UNIPROT por ser um banco de dados utilizado nesse trabalho e possuir o SWISS-PROT com informações curadas.

Após a classificação dos genes por função utilizando termos GO, foi necessário validar se os genes preditos realmente eram genes presentes em espécies de Leishmania. Para fazer essa validação foi comparado utilizando BLAST os genes preditos na *Leishmania braziliensis* com os genes anotados da *Leishmania panamensis*, [Llanes et al. 2015]. Foi escolhido a esta espécie para comparação pois pertence ao mesmo subgênero que a espécie alvo desse estudo. Nessa comparação foi utilizado como critério de similaridade sequências com mais de 40% de identidade, [Barnes 2007]. O e-value utilizado como parâmetro foi igual a 1e-5, [Xiong 2006]. Os demais parâmetros não foram alterados, mantendo os valores padrões implementados no algoritmo BLASTx, [Altschul et al. 1990].

Depois da busca dos genes anotados em um genoma de outra espécie do mesmo subgênero, foi feita outra comparação. Essa segunda comparação é realizada entre os genes anotados por esse trabalho, os anotados pelo NCBI e os anotados pelo TrytripDB. Para comparar as anotações foi feita a exportação das localizações gênicas de cada banco de dados para dois arquivos (.BED). Para efetuar essa comparação de coordenadas foi utilizado um programa do pacote BEDTools chamado *Intercept*, [Quinlan e Hall 2010]. Foi utilizado o parâmetro do Intercept que garante apenas overlap de 100%.

3.3 Predição de RNAs não codificantes

Foi executado o preditor de RNA não-codificante conhecido como structRNAFinder (`http://integrativebioinformatics.me/structrnafinder/`). Foi necessário instalar antes de executá-lo, as suas dependências que são: INFERNAL, RNAfolds e o BioPerl. Este algoritmo recebe como entrada o arquivo FASTA com todas as sequências do genoma. No comando do StructRNAFinder foi melhorado a performance do algoritmo utilizando o parâmetro para indicar a quantidade de núcleos ou processadores que poderão ser utilizados no processo limitando a três processadores.

Para efetuar uma predição sensível, foi realizada também a predição de RNA não codificante utilizando busca por similaridade. A base utilizada foi a NRDR (*The Non-coding RNA Database Resource*) [Paschoal et al. 2012]. A NRDR foi subdividida em três dataset's: o primeiro formado pelas sequências de tamanho superior a 1000 nucleotídeos, o segundo formado por sequências de tamanho entre 21 e 1000

nucleotídeos e o terceiro formado por sequências de tamanho menor que 21 nucleotídeos.

Foi executada uma busca por similaridade nos subdatasets utilizando dois algoritmos, o BLAT e o Bowtie. A execução dessa busca por similaridade com BLAT foi realizado utilizando a versão disponibilizada no seguinte endereço: `http://users.soe.ucsc.edu/~kent/src/blatSrc\protect\T1\textdollarver.zip`.

O comando do BLAT foi executado para todas as sequências do NRDR com tamanho superior a 21 nucleotídeos. Foram utilizados parâmetros para indicar ao algoritmo que todos os alinhamentos independente da sua pontuação e o seu grau de identidade, deveriam ser considerados. Também limitamos a quantidade de repetições por janela a 2253. O espaçamento entre as janelas foi configurado a 5 pares de bases. As sequências com tamanho inferior a 21 nucleotídeos, foram alinhadas contra o genoma utilizando o Bowtie. Inicialmente para executar este algoritmo é necessário formatar o dataset com o genoma de referência, nesse caso o *L. braziliensis*.

Após executar o *bowtie-build*(programa para formatar o banco de dados de referência), foi utilizado o programa bowtie-inspect para verificar se a formatação foi feita com sucesso. Com a base de dados formatada e verifica, pode-se então executar o algoritmo Bowtie de busca por similaridade. Para fazer o alinhamento entre o genoma e as sequências de RNA não codificante do NRDR menores que 21 nucleotídeos.

3.4 Estruturação dos dados gerados

Com os dados gerados, a próxima etapa é a estruturação desses dados em um modelo de dados relacional implementado no MySQL. Foi desenvolvido um modelo relacional apropriado para os dados gerados nesse projeto. Este desenvolvimento foi feito mediante a uma análise prévia dos dados, tipagem e dimensionamento dos mesmos.

3.5 O software de *frontend*

Após a etapa de análise e acomodação dos dados no modelo implementado no SGBD PostGreSQL, foi desenvolvido o *frontend*. Este software tem por objetivo permitir a visualização e download desses dados gerados. As funcionalidades desse software são demonstradas no diagrama de casos de uso da figura 3.2 foi desenvolvido.

Segue abaixo a descrição de cada requisito funcional e não funcional do sistema de *frontend* do LeishDB:

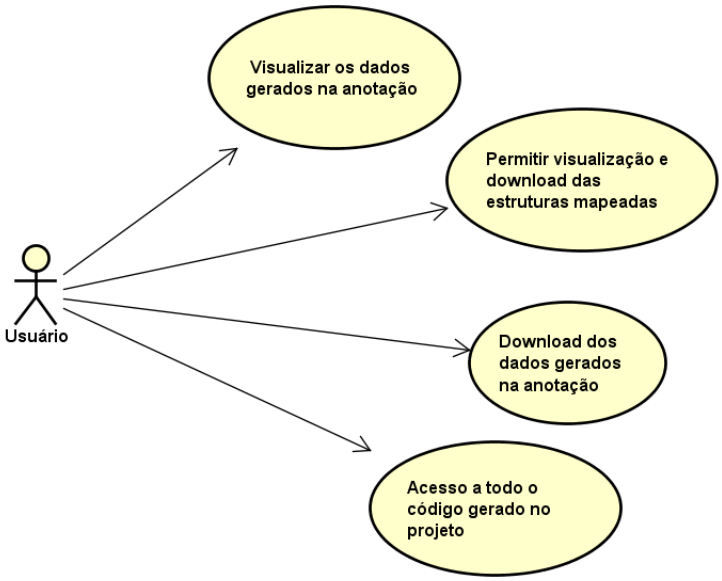

Figura 3.2: Diagrama de casos de uso do frontend do LeishDB. Fonte: Próprio Autor.

3.5.1 Requisitos funcionais

- **RF001 - Visualizar os dados gerados na anotação** Este requisito remete a função em que o usuário tem acesso as informações geradas no processo de anotação por meio de um *Genome Browser*. Este requisito é estável e essencial para este software pois é a sua principal funcionalidade. Possui também uma prioridade elevada devido a importância dessa função para o sistema.
- **RF002 - Permitir visualização e download das estruturas mapeadas** Este requisito remete a função do usuário fazer download das estruturas mapeadas no processo de anotação e visualizadas pelo *genome browser*. Este requisito é estável porém possui uma prioridade mediana.
- **RF003 - Download dos dados gerados na anotação** Este requisito remete a função em que o usuário pode efetuar o download de todos os dados gerados na anotação no formatos (.BED, .SAM, .CSV ou .SQL). Este requisito é estável e essencial para este software. Possui também uma prioridade elevada devido a importância dessa função para o sistema.
- **RF004 - Acesso a todo o código gerado no projeto** Este requisito remete a função em que o usuário tem acesso ao código do software desenvolvido para frontend. Este requisito é estável e essencial para este software pois é a sua principal funcionalidade. Possui também uma prioridade elevada devido a importância dessa função para o sistema.

3.5.2 Requisitos não-funcionais

- **RNF001 - Linguagem de Programação** Este requisito remete a necessidade do código do frontend ser desenvolvido em PHP devido a compatibilidade com o SGBD MySQL e demais linguagens de programação. Este requisito é estável e essencial de prioridade elevada.
- **RNF002 - SGBD** Este requisito remete a necessidade do banco de dados gerados nesse trabalho ser construído e gerenciado utilizando o SGBD MySQL. Este requisito é estável e essencial de prioridade elevada.
- **RNF002 - Servidor** Este requisito remete a necessidade do sistema ser compatível a execução em servidores Linux baseados no Debian. Este requisito é estável e essencial de prioridade elevada.
- **RNF003 - Sistema Web** Este requisito remete a necessidade do sistema de frontend ser um sistema web para acesso em navegadores de internet. Este requisito é estável e essencial de prioridade elevada.

3.5.3 Implementação do frontend e configuração do JBrowser

O frontend foi desenvolvido em scripts em PHP. Foram utilizadas boas práticas de programação para resultar em um código reutilizável. O template gráfico utilizado foi desenvolvido e disponibilizado pela página `http://themes.3rdwavemedia.com/`. Utilizando o template gráfico foi adicionado os códigos e conteúdos do portal e acoplado a ele um *genome browser*. O *genome browser* escolhido para utilização foi o JBrowser. Devido a arquitetura desse software foi necessário acoplar alguns scripts de Javascript ao portal.

Inicialmente foi feita a descompactação do código disponível no sítio `http://jbrowse.org/` do JBrowse 1.11.6. Após a descompressão foi feito o *deploy* desse código no servidor Apache 2.2. Depois da instalação do *genome browser* é necessário formatar o banco de dados para o modelo compatível com o JBrowser. O processo de formatação do modelo é representado graficamente na figura 3.3. O modelo compatível é o JSON e a formatação é feita baseada em arquivos .FASTA, .BED e .SAM. Para formatar a base de dados foi utilizado um script em PERL disponibilizado nas funções auxiliares do JBrowse.

O script em PERL adotou as sequências do arquivo .FASTA como referência. Após da definição das sequências de referências, foi feita a identificação das localizações presentes no arquivo .BED no .FASTA. Depois da identificação, foram buscadas informações adicionais sobre cada trecho no arquivo .SAM. Por fim todas as etapas acima foram armazenadas em scripts no formato JSON para consulta do JBrowse. A distribuição dos dados nos arquivos JSON são predefinidas pelo JBrowse e estão presentes na sua documentação.

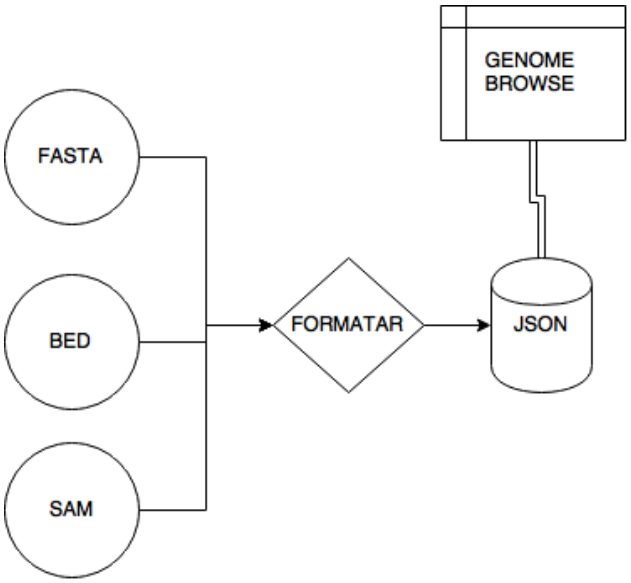

Figura 3.3: Diagrama de formatação dos dados para o JBrowse. Fonte: Próprio Autor.

Capítulo 4

Resultados

4.1 Dataset do genoma da *Leishmania braziliensis*

Inicialmente, foi necessário entender melhor as sequências presentes no dataset. Logo, verificou-se a existência de algo peculiar ao descrito na literatura em relação ao genoma da *L. braziliensis*. Foi comparado a quantidade de cromossomos presentes no dataset e a quantidade descrita por [Peacock et al. 2007]. Como pode ser observado na distribuição gráfica dos cromossomos por tamanho da sequência, demonstrado na Figura 4.1. A quantidade de cromossomos é igual a descrita em literatura, porém observa-se a ausência dos cromossomos 20 e 34.

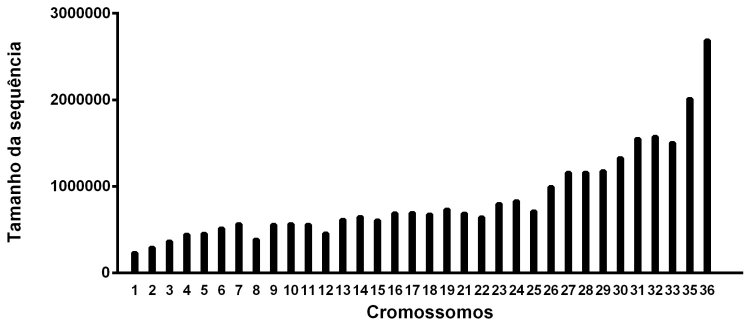

Figura 4.1: Distribuição gráfica dos cromossomos baixados da *Leishmania braziliensis* e o tamanho da sequência. Fonte: Próprio Autor.

O dataset do GeneDB (utilizado nesse trabalho), foi testado para aumentar a certeza de que as sequências presentes eram realmente de um organismo do gênero da *Leishmania*. Essa avaliação foi feita utilizando a correlação de Pearson entre o dataset e os dados das sequências de *L. infantum* do trabalho de (Wincker et al, 1996).

O gráfico demonstrado na Figura 4.2, demonstra a resultado da correlação com o P-value < 0.0001 e o $R^2 = 0,8001$. Essa correlação positiva demonstrou que existe um sinal que essa sequência seja do gênero *Leishmania*.

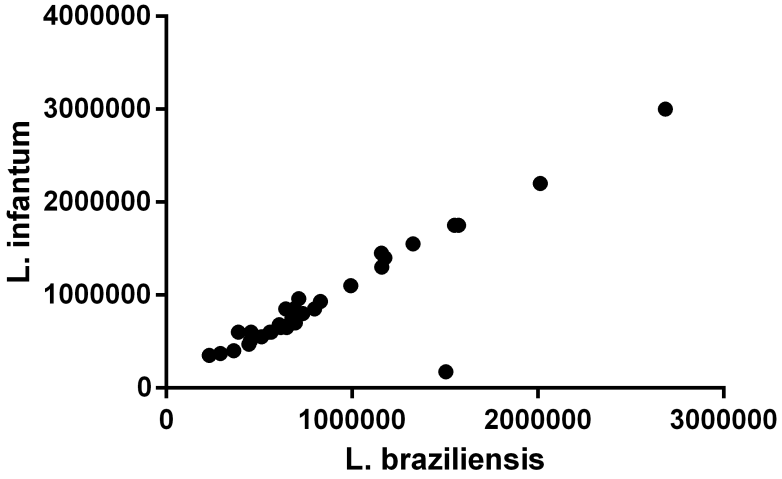

Figura 4.2: Correlação de Pearson entre o tamanho do genoma de *L. braziliensis* e *L. infantum*. Fonte: Próprio Autor.

4.2 Predição de genes codificantes de proteínas

O procedimento de predição gênica resultou em um total de 25539 ORF's identificadas no genoma da *L. braziliensis*, representada na Figura 4.3.A. Sendo 2266 (8.87%) dessas predições foram feitas pelo GENSCAN e 23284 (91.13%) pelo GLIMMER. Pode-se observar na Figura 4.3, a distribuição de quantidade de ORF's preditas pelos preditores em cada cromossomo. Por causa dessa alta taxa de falsos positivos, é necessário fazer uma filtragem dos dados preditos afim de validá-los.

Observando o gráfico da 4.3 e comparando com a distribuição por tamanho da 4.1, percebe-se uma proporção entre a quantidade de ORF's identificadas e o tamanho dos cromossomos da *L. braziliensis*. Fazendo a correlação de Pearson entre a quantidade de ORF's preditas pelo GLIMMER e o tamanho dos cromossomos foi encontrada uma correlação (Figura 4.4) quase perfeita com P-value < 0.0001 e o $R^2 = 0.9952$. A mesma correlação foi feita para as predições do GENSCAN e foi encontrada uma correlação(Figura 4.4) quase perfeita com P-value < 0.0001 e o $R^2 = 0.9737$.

Devido a correlação demonstrada na Figura 4.4, este foi o cromossomo com a maior quantidade de ORF's identificada em ambos os preditores. Pode-se observar

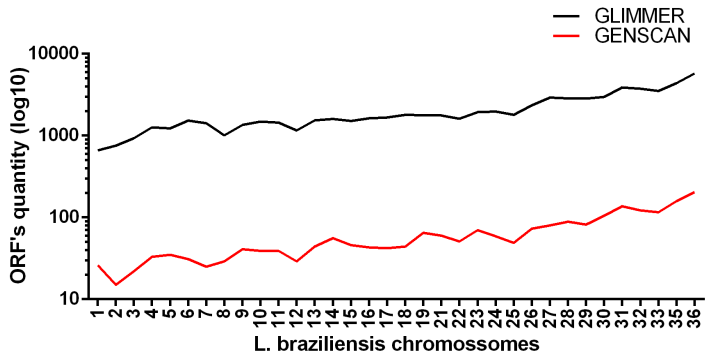

Figura 4.3: Distribuição de ORF's preditas em cada cromossomo do genoma da *L. braziliensis*. Fonte: Próprio Autor.

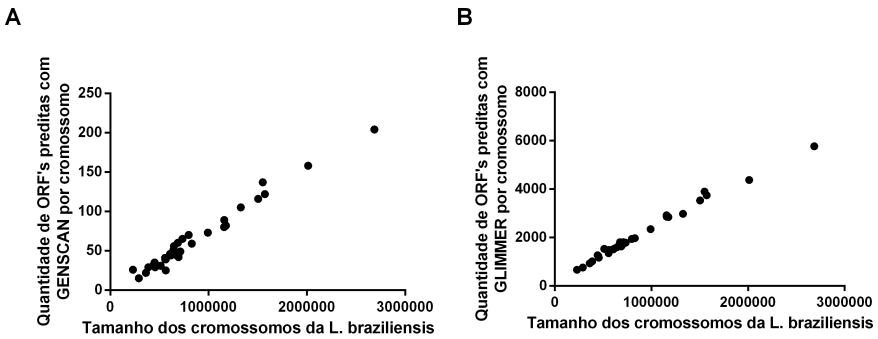

Figura 4.4: Gráfico de correlação de Pearson entre a quantidade de ORF's identificadas por cromossomo pelo GENSCAN (A) ou pelo GLIMMER (B) e o tamanho do cromossomo da *Leishmania braziliensis*. Fonte: Próprio Autor.

a localização das ORF's identificadas pelo GLIMMER e pelo GENSCAN nos cromossomos demonstrados na Figura 4.5.

Obtendo as localizações das ORF's identificadas por cada preditor foi possível caracterizá-los por similaridade afim de validar as predições e detectar a função dos genes preditos. Para executar a busca por similaridade foi utilizado o algoritmo BLAST implementado no software BLASTx entre os genes preditos e o banco de dados SWISS-PROT. Foram caracterizadas cerca de 4916 ORF's. Resultando em 2070 proteínas identificadas no genoma da *L. braziliensis*. Atualmente estão armazenadas no NCBI cerca de 661 proteínas associadas a genoma (nesta contabilização não foram levadas em consideração as proteínas hipotéticas e putativas de ambos os bancos).

Na Figura 4.6, ocorreu um aumento na quantidade de proteínas identificadas no genoma da *L. brazilisensis* . Foi executado o teste de Mann-Whitney após ter verificado

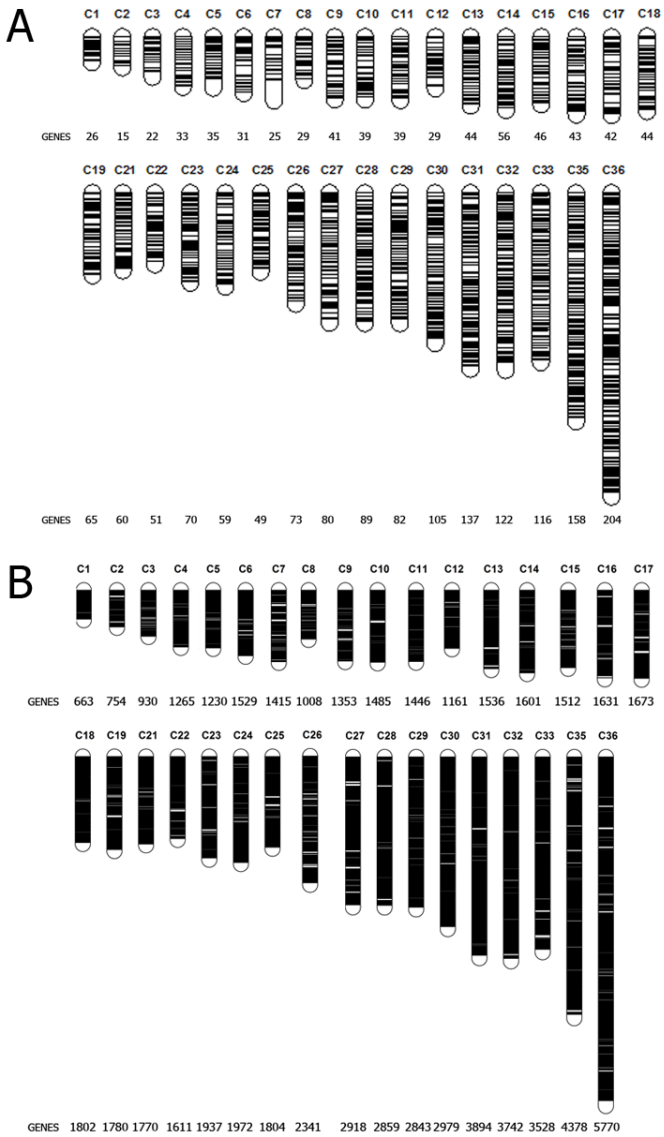

Figura 4.5: Representação gráfica da localização e o total de ORF's preditas pelos preditores GENSCAN (A) e GLIMMER (B) por cromossomo. Fonte: Próprio Autor.

pelo teste D'Agostino & Pearson que os dados não seguiam uma distribuição Gaussiana. Pode-se então sugerir um aumento significativamente estatístico com o *P-value* = 0.0003. Após o teste estatístico, foram comparadas as anotações dessa espécie do NCBI e a executada nesse trabalho. Essa comparação resultou na confirmação de: 2970 genes que eram classificados pelo NCBI como hipotéticos ou putativos tiveram as suas proteínas identificadas e foram identificados 1223 novos genes com as suas proteínas identificadas. O gráfico com as estatísticas gerais da comparação dos genes está representado na figura 4.7.

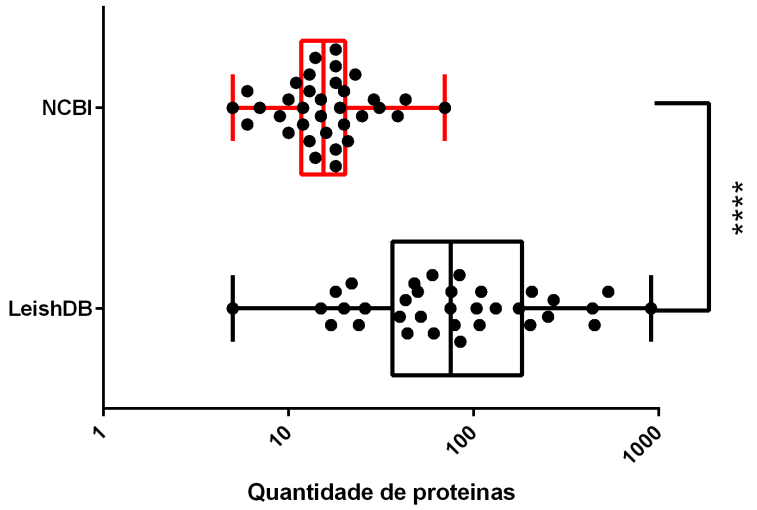

Figura 4.6: Quantidade de proteínas preditas por cromossomo. Fonte: Próprio Autor.

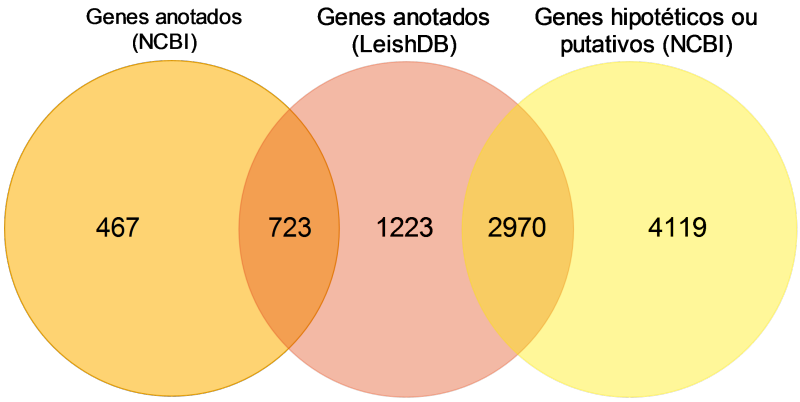

Figura 4.7: Comparação entre a anotação atual da *L. braziliensis* disponibilizada pelo NCBI e a anotação desse trabalho. Fonte: Próprio Autor.

Após a comparação com a anotação do NCBI, foi feita uma comparação também com a anotação do TritrypDB. Essa comparação resultou na confirmação de: 1951 genes que eram classificados como hipotéticos ou putativos pela anotação do TritrypDB tiveram suas proteínas identificadas e foram identificadas 2509 novos genes com as suas proteínas identificadas. O gráfico que representa essa comparação, está representado na figura 4.8.

As proteínas identificadas por este trabalho, tiveram as suas funções identificadas através da busca por termos *Genes Onthology* (GO). Utilizando a ferramenta web AmiGO, foram identificadas por termos GO cerca de 201 proteínas de 2070

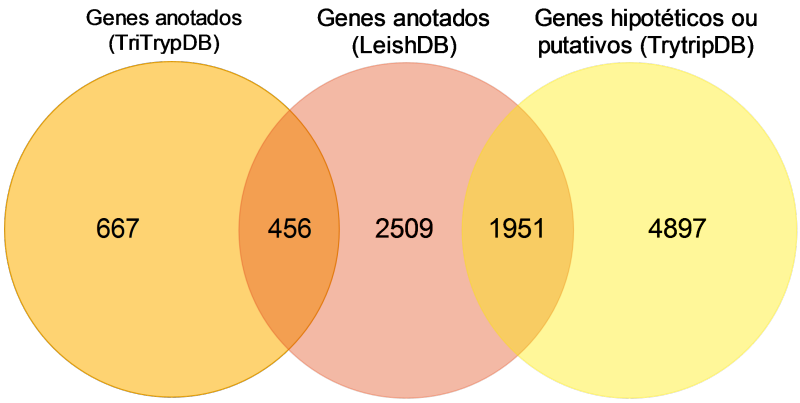

Figura 4.8: Comparação entre a anotação atual da *L. braziliensis* disponibilizada pelo TriTrypDB e a anotação desse trabalho. Fonte: Próprio Autor.

proteínas identificadas. Foram classificada em três domínios (*clusters*): componente celular (Figura 4.9), função molecular (Figura 4.10) e processo biológico (Figura 4.10). Foram utilizados para classificar as 201 proteínas, cerca de 945 termos GO que foram condensados pelo AmiGO em 121 termos.

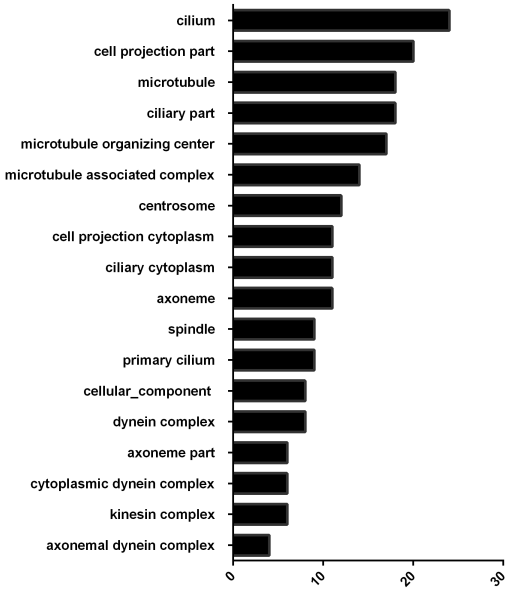

Figura 4.9: Distribuição de proteínas por domínio de componente celular classificado por termo GO. Fonte: Próprio Autor.

Após a identificação das funções das proteínas categorizadas, foi necessário validar

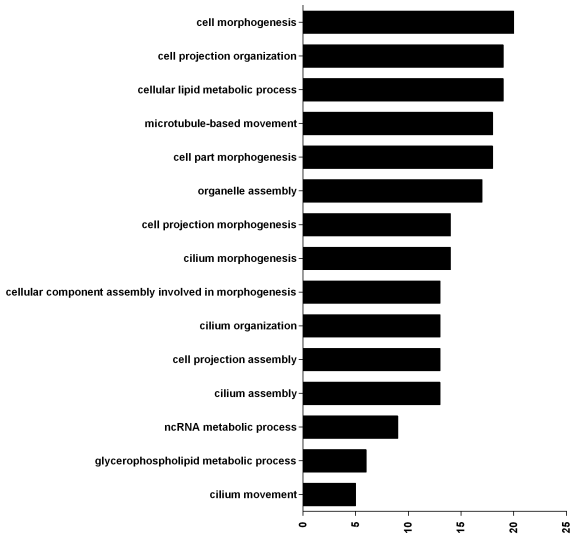

Figura 4.10: Distribuição de proteínas por domínio de processo biológico classificado por termo GO. Fonte: Próprio Autor.

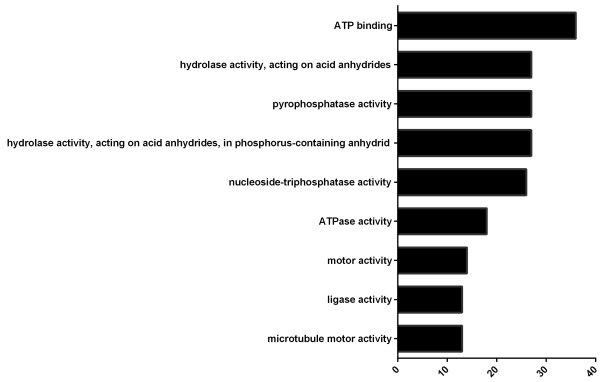

Figura 4.11: Distribuição de proteínas por domínio de função molecular classificado por termo GO. Fonte: Próprio Autor.

os genes encontrados. Essa validação consiste na verificação da existência de genes preditos na *L. braziliensis* em outra espécie do mesmo subgênero. O resultado do BLAST feito entre os genes preditos da *L. braziliensis* e a anotação disponível da *L. panamensis* garantiu a presença de 94.75% dos genes preditos. Ou seja, dos genes preditos (4916) para *L. braziliensis,* 4658 (94.75%) tiveram similaridade com genes da *L. panamensis*. Esses percentuais estão descritos graficamente na figura 4.12.

Após a evidência de aproximadamente 95% dos genes preditos no genoma da *L. braziliensis* estarem presentes também no genoma da *L. panamensis*. Nós fizemos a contagem de cópias de genes mapeados. Essa lista de cópias de genes foi ranqueada

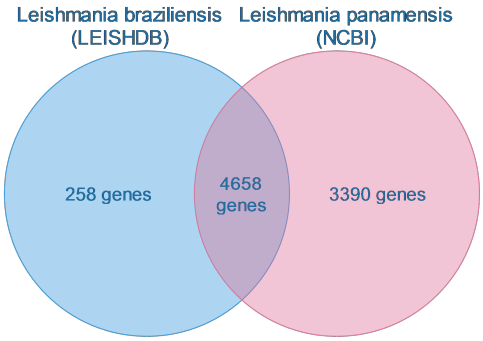

Figura 4.12: Diagrama de comparação entre a *L. braziliensis* e a *L. panamensis*. Fonte: Próprio Autor.

pelo número de cópias e representada na Tabela 4.2.

Número de Cópias	Localização (chr)	SWISSPROT ID	Descrição
849	16	P20481	Buccalin
484	22	P11976	Prestalk protein
410	15	Q9W596	Microtubule-associated
47	4	P12027	Polysialoglycoprotein (PSGP)
22	10	Q00689	Leishmanolysin
11	31	O74327	Vacuolar amino acid transporter 5
8	16	O00874	DNA polymerase alpha
7	30	P12076	Heat shock 70-related
7	23	Q9MBF8	Dynein-1-beta

Tabela 4.1: Ranking de cópias protêicas presentes na anotação do LeishDB. Fonte: Próprio autor.

4.3 Predição de RNAs não codificantes

Utilizando o StructRNAFinder (`http://integrativebioinformatics.me/structrnafinder/`) foi feita a predição de RNAs não codificantes e a classificação por família. Foram identificadas cerca de 735 RNAs, a classificação desses RNAs estão demonstradas na Figura 4.13.

Os 735 RNAs foram classificados em 9 famílias sendo elas: IRES (10), leader (1), lncRNA (14), miRNA (421), rRNA (11), snRNA (2), snoRNA (147), sRNA (16), tRNA (6) e outras famílias (37). A média de hits por cromossomo foi cerca de 21,62. Com o mapeamento dos RNAnc no genoma, pode-se construir um mapa

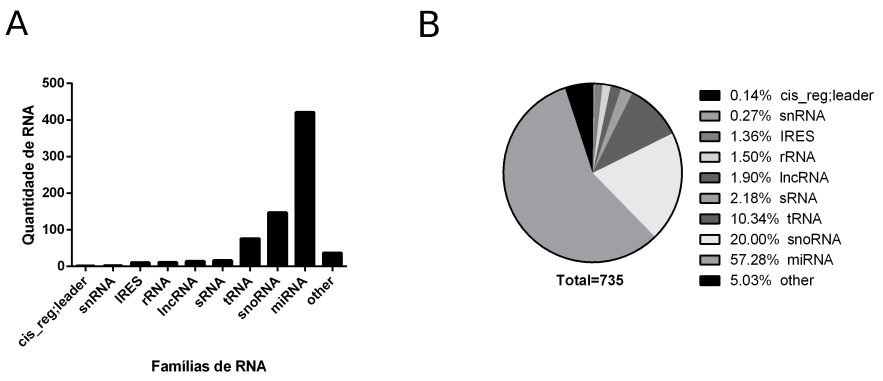

Figura 4.13: Distribuição de RNAs não codificantes preditos por família, absoluta (A) e relativa (B). Fonte: Próprio Autor.

genômico circular com todos os genes codificantes confirmados e mapeados e os RNAnc também mapeados e confirmados.

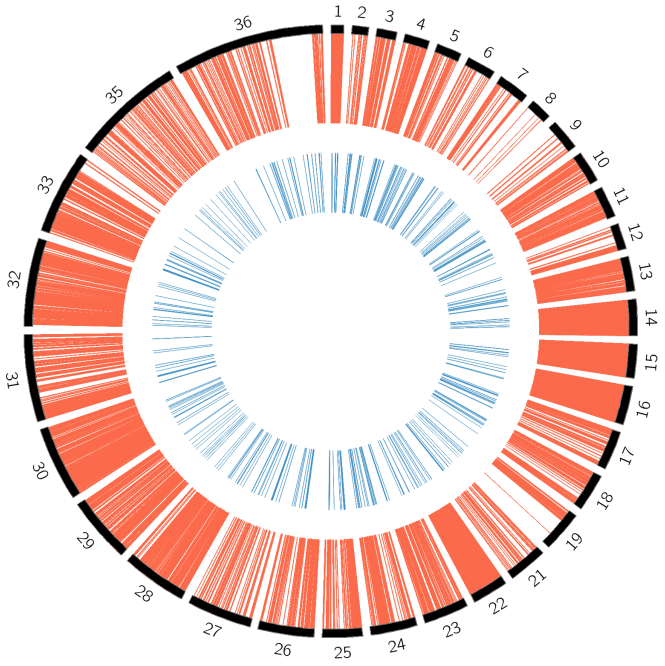

Figura 4.14: Mapa genômico das estruturas preditas e confirmadas nesse trabalho. Em vermelho os genes codificantes e em azul os ncRNAS. Fonte: Próprio Autor.

No mapa genômico da figura 4.14 gerado apartir dos dados desse projeto, pode-se perceber algumas áreas ricas em RNAnc. Essas regiões chamaram atenção e foi realizada a contagem de ncRNA mapeados em cada cromossomo.

Analisando a distribuição da figura 4.15 há um alto índice de RNA não codificante

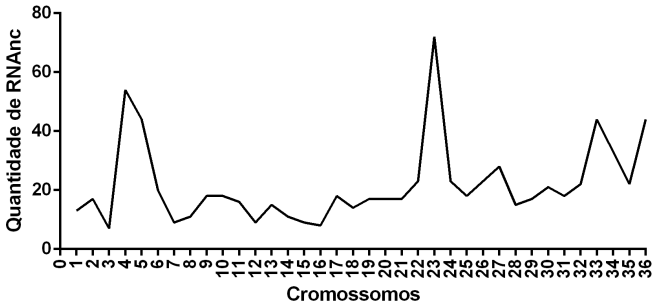

Figura 4.15: Distribuição de ncRNA por cromossomo do genoma da *Leishmania braziliensis*. Fonte: Próprio Autor.

nos cromossomos: 4,5 e 23. Baseado nessa característica e utilizando a dados literários disponibilizados na base de dados do IPA (*Ingenuity Pathway Analysis*), analisamos as interações protêicas dessas regiões. Essa análise resultou na detecção de duas regiões que possuem um alto número de interações, o cromossomo 5 e o 23.

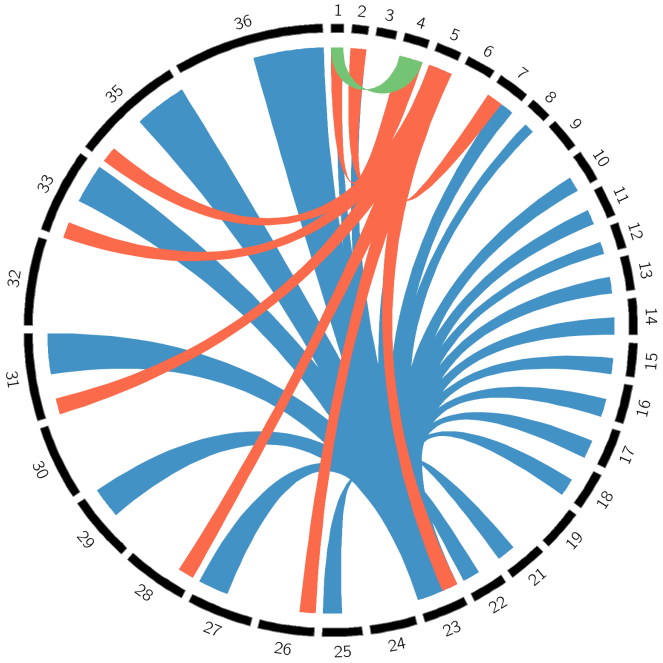

Figura 4.16: Interações protêicas dos cromossomos 4 (verde), 5 (vermelho) e 23 (azul) da *Leishmania braziliensis*. Fonte: Próprio Autor.

Todos os dados gerados foram armazenados no modelo de banco de dados representado na Figura 4.17. O modelo é composto por cinco entidades sendo elas: *chromossomes*, *genes*, *ncrna*, *proteins* e *gene_ontology*.

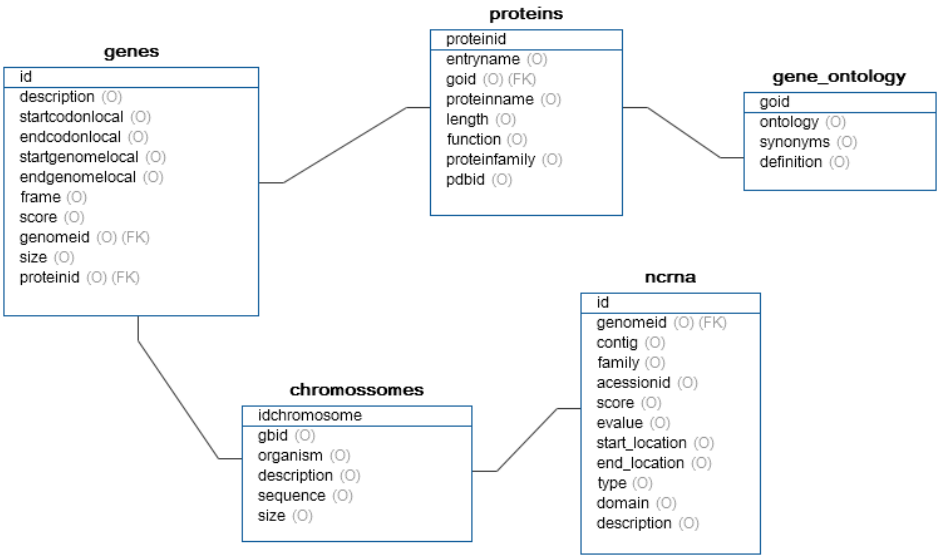

Figura 4.17: Representação gráfica do modelo relacional do banco de dados do LeishDB. Fonte: Próprio Autor.

4.4 O software de *frontend*

Foi implementado um website com os dados gerados na anotação disponíveis para download. O endereço desse website é: `http://www.leishdb.com`. A tela inicial do site está representada na figura 4.18. Nessa tela o usuário pode efetuar buscas na base LeishDB de componentes presentes nessa anotação. Para efetuar a busca simples o usuário deve digitar no campo de texto: o nome da proteína, o SWISS-PROT ID ou o Gene ID do LeishDB.

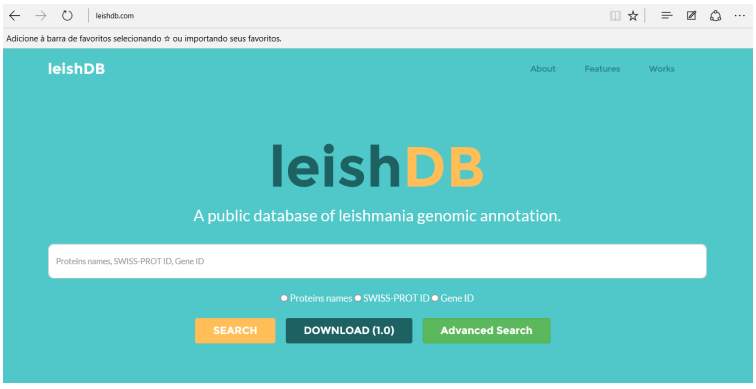

Figura 4.18: Tela inicial do sistema de *frontend*. Fonte: Próprio Autor.

O outro tipo de busca é a busca avançada. Essa busca permite que o usuário busque componentes pelo número do cromossomo da *L. braziliensis* ou pelo filtro de coordenada. Após a escolha que atenda os critérios de componentes alvo, o usuário

clica no botão "*Search*". O sistema irá efetuar a busca no banco de dados e em seguida retornará para a tela de resultados com os componentes encontrados.

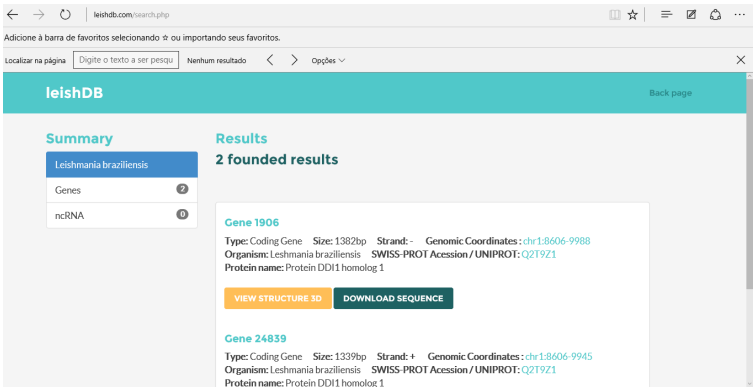

Figura 4.19: Tela de resultados da busca no sistema de *frontend*. Fonte: Próprio Autor.

Na figura 4.19, está representada a tela de resultados de uma busca efetuada no banco de dados LeishDB. Sua estrutura é simplificada para facilitar a visualização dos resultados. Ao lado esquerdo da tela está presente um resumo dos componentes encontrados na busca. Em seguida são listados todos os componentes encontrados, suas informações e opções. Existem na lista de registros links para o JBrowser que permitem visualizar o componente no cromossomo. Na figura 4.20, pode ser visualizado a representação gráfica da localização do Gene 25274 e a sequência desse componente.

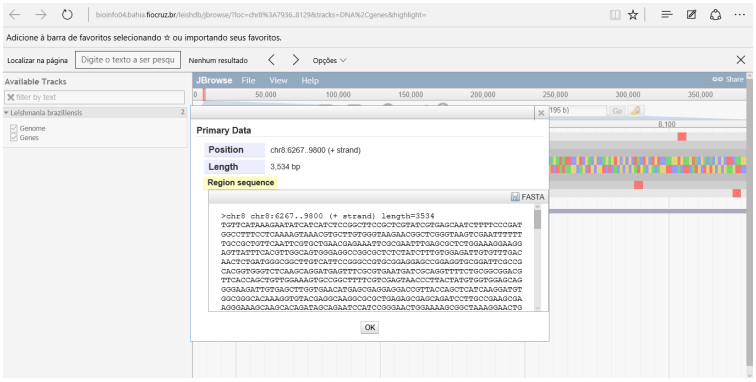

Figura 4.20: Tela do JBrowser representando o Gene 25274. Fonte: Próprio Autor.

No LeishDB é possível baixar o FASTA de todos os componentes utilizando a opção "*Download Sequence*". Os genes que tiveram a sua tradução anotada neste trabalho possuem a disponibilidade de visualização da estrutura 3D da proteína.

Capítulo 5

Discussões

Uma das observações de acordo com os resultados obtidos, foi a ausência dos cromossomos 20 e 34 no dataset do GeneDB da *L. braziliensis*. O *dataset* disponível no NCBI e no TrytripDB também não apresentavam esses cromossomos. Isto se deve a fusão dos cromossomos 20 e 34, representada pelo cromossomo 36,[Laurentino et al. 2004]. Essa fusão resultou no cromossomo 36 o maior tamanho em pares de bases em comparação aos demais. O que acaba sendo um ponto de dispersão nos gráficos de correlação entre cromossomos apresentados na sessão de resultados. O trabalho de [Laurentino et al. 2004], realizou uma anotação nesse organismo. Porém, existiu nesse trabalho a presença de muitos genes hipotéticos ou putativos. Analisando a metodologia aplicada e a revisão de [Yandell e Ence 2012], esta presença ocorreu devido a utilização de bases de dados de proteínas não curadas.

Realizamos uma confirmação para verificar se o dataset baixado era realmente de um organismo da espécie *Leishmania*. O resultado da correlação confirma que o dataset do GeneDB do genoma da *L. braziliensis* tem uma correlação quase perfeita com o genoma do *L. infantum* descrito na literatura. Pode-se afirmar então que esse dataset é composto por sequências de um organismo do gênero *Leishmania*. Como ele também já foi utilizado na literatura no trabalho de [Peacock et al. 2007], temos segurança sobre os dados utilizados.

Na fase de predição dos genes codificantes, percebemos que o GLIMMER identificou muito mais ORF's que o GENSCAN. Isso ocorreu devido a diferença da complexidade entre o genoma de eucariotos e de procariotos, [Aggarwal et al. 2003]. Como o procarioto tem a sua estrutura gênica bem delimitada, o preditor para procarioco é sensível a este tipo de estrutura e tende a encontrar um valor grande de ORF's. Baseados na comparação entre as distribuições das ORF's nós testamos a hipótese de se existe uma correlação entre a quantidade de ORF's e o tamanho dos cromossomos. Após testar com correlação estatística, pode-se dizer que existe uma correlação entre a quantidade de ORF's preditas e o tamanho dos cromossomos.

As estruturas codificantes e não codificantes preditas foram mapeadas no genoma. Após a predição e mapeamento nós efetuamos o cruzamento desses dados utilizando

inicialmente a comparação dos mapeamentos. Percebemos então que as regiões (cromossomo 5 e 23) que tiveram maior quantidade de genes confirmados nesse trabalho, também tiveram uma maior quantidade de rna não codificante preditos. Observando este fato, levantamos a hipótese de que exista algum relacionamento entre genes e rna não codificantes como os descritos no trabalho de [Kelada et al. 2013]. Para tentar entender esse fato, recorremos a literatura científica utilizando o IPA para verificar as interações descritas entre as proteínas caracterizadas. O resultado dessa análise utilizando os dados do IPA evidênciou que os cromossomos que possuem uma maior quantidade de rna não codificantes possuem uma larga rede de interação protêica entre os demais cromossomos.

Após a predição dos genes, foram realizadas as comparações entre a anotação do LeishDB (realizada nesse trabalho) e as anotações de outros bancos de dados (NCBI e TritrypDB). Essa comparação demonstrou que na anotação desse trabalho, foram confirmados 2970 genes hipotéticos da anotação NCBI e 1951 genes hipotéticos da anotação TriTrypDB. Encontramos também cerca de 2509 novos genes, que não estavam descritos no NCBI e 2034 novos genes não descritos no TritrypDB. Percebemos que em ambas comparações não foram encontrados por esse trabalho todos os genes descritos e caracterizados nas anotações desses outros bancos. Essa disparidade ocorreu devido a utilização de uma base de dados de proteína para caracterização dos genes. Pois todos foram preditos, porém na etapa de caracterização das ORF's, foram filtrados. Este fato ocorreu por não possuirem uma similaridade acima do *cutoff* com alguma proteína referência.

Com base nesses dados, esse trabalho aumentou o conhecimento sobre o genoma da *L. braziliensis*, caracterizando genes hipotéticos e predizendo e confirmando genes desconhecidos anteriormente. Os RNA's não codificantes na *Leishmania braziliensis* são pouco estudados e este estudo demonstrou uma potencial interação desses elementos na interação protêica. Porém, é necessário no futuro estudar com mais afinco as interações protêicas identificadas pelo IPA e quantificadas nesse trabalho e efetuar testes da sua existência *in vivo*. Já que a função de alguns rna's não codificantes já foi definida na literatura como catalítica em processos nucleares,[Kelada et al. 2013]. Talvez o entendimento dessas interações permitar a intervenção de processos vitais para os protozoários dessa espécie possibilitando a cura ou novos métodos de tratamento desta patologia.

Fazendo uma comparação entre o estado da arte de metodologias de anotação e a utilizada nesse trabalho pode-se perceber que: a quantidade de genes anotados e confirmados foram maiores que a anotação disponível NCBI, [Laurentino et al. 2004]. Esta diferença deve-se basicamente a dois aspectos metodológicos: a qualidade da base de proteínas utilizada para identificação de função e a possibilidade de controle total dos parâmetros utilizados nesse trabalho. Muitos *pipelines*, não permitem a personalização de todos os parâmetros de uma forma fácil. Isso ocorre devido a automatização dos processos por meio de interfaces gráficas fixas. Por este motivo, nesse trabalho optamos pela execução de cada algoritmo e etapa do processo de anotação, individualmente.

Capítulo 6

Considerações Finais

Foi perceptível ao longo desse trabalho a carência de estudos dessa natureza para a *Leishmania braziliensis*. Por exemplo, no pubmed existem poucos estudos abordando a anotação desse parasito, que é o principal responsável pelo LTA no Brasil. O difícil acesso aos dados anotados em outros projetos e a variabilidade do formatos dos arquivos dificulta o uso. Para evitar esse problema, tentamos facilitar ao máximo o acesso aos dados gerados, seja: por sistema de frontend, genome browser ou baixando os arquivos. Ficou claro durante o processo de anotação que é necessário unificar os formatos de arquivos para anotação afim de facilitar o *pipelines*.

Com os resultados apresentados, foi salientada a importância da anotação gênica para descoberta de regiões e o mapeamento das conhecidas. Desta forma, outros projetos que utilizem o genoma da *L. braziliensis* podem localizar regiões alvo com maior facilidade. Pode-se também garantir um maior conhecimento gênico das estruturas do genoma dessa espécie se tratando de localizações e expressões gênicas. A importância do mapeamento dos RNA's não codificantes foi demonstrado, sugerindo a participação dessas moléculas em iterações protêicas.

Porém este projeto é o início de uma série de outros questionamentos que podem ser respondidos após o processo de anotação gênica. Pretende-se no futuro utilizar as estruturas identificadas para efetuar uma comparação entre as principais espécies de *Leishmania* no Brasil. Afim de encontrar genes e estruturas não codificantes em regiões conservadas entre as espécies. Após a comparação será possível buscar por exemplo regiões promotoras de resistência a drogas utilizadas no tratamento. A identificação dessas regiões, permitiriam ao médico optar por um tratamento mais eficaz para aquele paciente.

Referências Bibliográficas

[Aggarwal et al. 2003] Aggarwal, G., Worthey, E. a., McDonagh, P. D., e Myler, P. J. (2003). Importing statistical measures into Artemis enhances gene identification in the Leishmania genome project. *BMC bioinformatics*, 4:23.

[Almeida et al. 2004] Almeida, L. G. P., Paixão, R., Souza, R. C., da Costa, G. C., Barrientos, F. J. A., dos Santos, M. T., de Almeida, D. F., e Vasconcelos, A. T. R. (2004). A System for Automated Bacterial (genome) Integrated Annotation–SABIA. *Bioinformatics (Oxford, England)*, 20(16):2832–3.

[Altschul et al. 1990] Altschul, S., Gish, W., e Miller, W. (1990). Basic Local Alignment Search Tool. *J Mol Biol.*, 215(3):403–410.

[Barnes 2007] Barnes, M. R. (2007). *Bioinformatics for Geneticists*. Number 2.

[Benson et al. 2014] Benson, D. a., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., e Sayers, E. W. (2014). GenBank. *Nucleic acids research*, 42(Database issue):D32–7.

[Besemer et al. 2001] Besemer, J., Lomsadze, A., e Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic acids research*, 29:2607–2618.

[Brent 2005] Brent, M. R. (2005). Genome annotation past , present , and future : How to define an ORF at each locus Genome annotation past , present , and future : How to define an ORF at each locus. pp. 1777–1786.

[Burge e Karlin 1997] Burge, C. e Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology*, 268:78–94.

[Burge et al. 2013] Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P., e Bateman, A. (2013). Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research*, 41(November 2012):226–232.

[Cantarel et al. 2008] Cantarel, B. L., Korf, I., e Robb, S. M. C. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome ...*, pp. 188–196.

[Carbon et al. 2009] Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Lomax, J., Mungall, C., Hitz, B., Balakrishnan, R., Dolan, M., Wood, V., Hong, E., e Gaudet, P. (2009). AmiGO: Online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289.

[Carver et al. 2012] Carver, T., Harris, S. R., Berriman, M., Parkhill, J., e McQuillan, J. a. (2012). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, 28(4):464–469.

[Delcher et al. 2007] Delcher, A. L., Bratke, K. a., Powers, E. C., e Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, 23(6):673–679.

[Dias Correia e a.a Dias Correia 2007] Dias Correia, J. e a.a Dias Correia (2007). Funcionalidades dos RNA não codificantes (ncRNA) e pequenos RNA reguladores , nos mamíferos. *REDVET: Revista electrónica de Veterinaria*, VIII(1695-7504):1–22.

[Eddy e Durbin 1994] Eddy, S. R. e Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic acids research*, 22(11):2079–2088.

[Elmasri e Navathe 2011] Elmasri, R. e Navathe, S. B. (2011). *Fundamentals of database systems*. Pearson Education, 6 edition.

[Ewing e Green 1998] Ewing, B. e Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research*, (206):175–185.

[Fern et al. 2015] Fern, M., Galperin, M. Y., e Rigden, D. J. (2015). The 2015 Nucleic Acids Research Database Issue and. 43:1–5.

[Gontijo e Melo 2004] Gontijo, C. M. F. e Melo, M. N. (2004). Leishmaniose visceral no Brasil: quadro atual, desafios e perspectivas.

[Guigó et al. 1992] Guigó, R., Knudsen, S., Drake, N., e Smith, T. (1992). Prediction of gene structure. *Journal of molecular biology*, 226:141–157.

[Johnson et al. 2008] Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., e Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic acids research*, 36(Web Server issue):5–9.

[Karolchik et al. 2009] Karolchik, D., Kent, W. J., e Angie S. Hinrichs (2009). *The UCSC Genome Browser*. Curr Protoc Bioinformatics.

[Kelada et al. 2013] Kelada, S., Sethupathy, P., Okoye, I. S., Kistasis, E., Czieso, S., White, S. D., Chou, D., Martens, C., Ricklefs, S. M., Virtaneva, K., Sturdevant, D. E., Porcella, S. F., Belkaid, Y., Wynn, T. a., e Wilson, M. S. (2013). miR-182 and miR-10a Are Key Regulators of Treg Specialisation and Stability during Schistosome and Leishmania-associated Inflammation. *PLoS Pathogens*, 9(6).

[Kelley et al. 2012] Kelley, D. R., Liu, B., Delcher, A. L., Pop, M., e Salzberg, S. L. (2012). Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Research*, 40.

[Kent 2002] Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664.

[Kent e Zahler 2000] Kent, W. J. e Zahler, a. M. (2000). The intronerator: exploring introns and alternative splicing in Caenorhabditis elegans. *Nucleic acids research*, 28(1):91–93.

[Korf 2004] Korf, I. (2004). Gene finding in novel genomes. *BMC bioinformatics*, 5:59.

[Langmead et al. 2009] Langmead, B., Trapnell, C., Pop, M., e Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25.

[Larsen e Krogh 2003] Larsen, T. S. e Krogh, A. (2003). EasyGene - a prokaryotic gene finder that ranks ORFs by statistical significance.

[Laurentino et al. 2004] Laurentino, E. C., Ruiz, J. C., Fazelinia, G., Myler, P. J., Degrave, W., Alves-Ferreira, M., Ribeiro, J. M. C., e Cruz, A. K. (2004). A survey of Leishmania braziliensis genome by shotgun sequencing. *Molecular and biochemical parasitology*, 137(1):81–6.

[Li et al. 2014] Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., e Cui, Q. (2014). HMDD v2.0: A database for experimentally supported human microRNA and disease associations. *Nucleic Acids Research*, 42(D1):1070–1074.

[Llanes et al. 2015] Llanes, A., Restrepo, C. M., Vecchio, G. D., Anguizola, F. J., e Lleonart, R. (2015). The genome of Leishmania panamensis: insights into genomics of the L. (Viannia) subgenus. *Scientific Reports*, 5:8550.

[Majoros et al. 2004] Majoros, W. H., Pertea, M., e Salzberg, S. L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics*, 20(16):2878–2879.

[Maria et al. 2010] Maria, D., Aquino, C. D., Silva, A., Costa, A., e Barral, A. P. (2010). Avaliação imunológica da intradermorreação de Montenegro. 11(2):9–13.

[Marsden e Nonata 1975] Marsden, P. D. e Nonata, R. (1975). Mucocutaneous leishmaniasis - a review of clinical aspects.

[McQuilton et al. 2012] McQuilton, P., St Pierre, S. E., e Thurmond, J. (2012). FlyBase 101–the basics of navigating FlyBase. *Nucleic acids research*, 40(Database issue):D706–14.

[Ministério da Saúde Secretaria de Vigilância em Saúde 2007] Ministério da Saúde Secretaria de Vigilância em Saúde (2007). *Manual de vigilância da Leishmaniose Tegumentar Americana*. Editora do Ministério da Saúde, 2º edição edition.

[Moreira 1895] Moreira, J. (1895). Distribuição geográfica. *Gazeta Médica*.

[Nawrocki e Eddy 2013] Nawrocki, E. P. e Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935.

[Paschoal et al. 2012] Paschoal, A. R., Maracaja-Coutinho, V., Setubal, J. a. C., Simões, Z. L. P., Verjovski-Almeida, S., e Durham, A. M. (2012). Non-coding transcription characterization and annotation: A guide and web resource for non-coding RNA databases. *RNA Biology*, 9(3):274–282.

[Peacock et al. 2007] Peacock, C. S., Seeger, K., Harris, D., Murphy, L., Ruiz, J. C., Quail, M. a., Peters, N., Adlem, E., Tivey, A., Aslett, M., Kerhornou, A., Ivens, A., Fraser, A., Rajandream, M.-A., Carver, T., Norbertczak, H., Chillingworth, T., Hance, Z., Jagels, K., Moule, S., Ormond, D., Rutter, S., Squares, R., Whitehead, S., Rabbinowitsch, E., Arrowsmith, C., White, B., Thurston, S., Bringaud, F., Baldauf, S. L., Faulconbridge, A., Jeffares, D., Depledge, D. P., Oyola, S. O., Hilley, J. D., Brito, L. O., Tosi, L. R. O., Barrell, B., Cruz, A. K., Mottram, J. C., Smith, D. F., e Berriman, M. (2007). Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nature genetics*, 39(7):839–847.

[Pearson 2014] Pearson, W. R. (2014). An Introduction to Sequence Similarity (Homology) Searching. pp. 1–9.

[Penna 1934] Penna, H. (1934). Leishmaniose visceral no Brasil. *Bras Méd*, 48:949–50.

[Quinlan e Hall 2010] Quinlan, A. R. e Hall, I. M. (2010). BEDTools : a flexible suite of utilities for comparing genomic features. 26(6):841–842.

[Rutherford et al. 2000] Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. a., e Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics (Oxford, England)*, 16(10):944–945.

[Salzberg et al. 1998] Salzberg, S. L., Deicher, A. L., Kasif, S., e White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2):544–548.

[Skinner et al. 2009] Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J., e Holmes, I. H. (2009). JBrowse : A next-generation genome browser. *Genome research*, 19:1630–1638.

[Solovyev et al. 1994] Solovyev, V. V., Salamov, A. A., e Lawrenc, C. B. (1994). The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames.

[Stanke e Waack 2003] Stanke, M. e Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(Suppl 2):ii215–ii225.

[Vieira 2008] Vieira, S. (2008). *Introdução à Bioestatística*. Elsevier, Rio de Janeiro, 4° edition.

[Wallis et al. 2012] Wallis, L., Souza, F., Cristina, A., Botelho, C., Vilas, S., e Souza, T. (2012). Comparative analysis of the geographic distribution of the histopathological spectrum and Leishmania species of American. *An Bras Dermatol*, 87(3):369–374.

[Xiong 2006] Xiong, J. (2006). *Essential Bioinformatics*.

[Yandell e Ence 2012] Yandell, M. e Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature reviews. Genetics*, 13(5):329–42.

[Zhang 2002] Zhang, M. Q. (2002). Computational prediction of eukaryotic protein-coding genes. *Nature reviews. Genetics*, 3(9):698–709.

[Zhu et al. 2010] Zhu, W., Lomsadze, A., e Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic acids research*, 38(12):e132.