Guilherme Augusto Maia

FERRAMENTA INTEGRADA PARA ANOTAÇÃO DE PROTEÍNAS HIPOTÉTICAS: ESTUDO DE CASO UTILIZANDO ANÁLISES PROTEOGENÔMICAS EM *Trypanosoma rangeli*

Dissertação submetida ao Programa de Biotecnologia e Biociências da Universidade Federal de Santa Catarina para a obtenção do título de Mestre em Biotecnologia e Biociências sob orientação do Professor Dr. Glauber Wagner.

Florianópolis
2019

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Maia, Guilherme Augusto
Ferramenta integrada para anotação de proteínas hipotéticas : estudo de caso utilizando análises proteogenômicas em Trypanosoma rangeli / Guilherme Augusto Maia ; orientador, Dr. Glauber Wagner , 2019.
79 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências Biológicas, Programa de Pós-Graduação em Biotecnologia e Biociências, Florianópolis, 2019.

Inclui referências.

1. Biotecnologia e Biociências. 2. Bioinformática. 3. Trypanosomatidae. 4. Proteogenômica. 5. Porteínas hipotéticas. I. , Dr. Glauber Wagner. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Biotecnologia e Biociências. III. Título.

"Ferramenta integrada para anotação de proteínas hipotéticas: estudo de caso utilizando análises proteogenômicas em *Trypanosoma rangeli*"

Por

Guilherme Augusto Maia

Dissertação julgada e aprovada em sua forma final pelos membros titulares da Banca Examinadora (019/2019/PPGBTC) do Programa de Pós-Graduação em Biotecnologia e Biociências - UFSC.

Prof. Dr. Glauber Wagner
Coordenador do Programa de Pós-Graduação em Biotecnologia e Biociências

Banca examinadora:

Dr. Glauber Wagner (Universidade Federal de Santa Catarina)
Orientador

Dr. Guilherme de Toledo e Silva (Universidade Federal de Santa Catarina)

Dr. Daniel Santos Mansur (Universidade Federal de Santa Catarina)

Florianópolis, 31 de maio de 2019.

AGRADECIMENTOS

À minha família, meu pai Edisio, minha mãe Rosana e minha irmã Juliana, que me apoiaram, incentivaram e sempre estiveram do meu lado. Com certeza, todo o carinho de vocês tornou essa jornada muito mais fácil. Amo muito vocês.

Aos amigos e amigas que estiveram juntos nessa caminhada de dois anos, os quais quebram as barreiras físicas dos laboratórios e salas de aula do MIP: Caibe, por ser uma pessoa muito gente boa, super *good vibes*, também por toda a camaradagem e amizade desde o início do nosso mestrado; Leandra, minha amiga desde os tempos de graduação, por ser sempre uma pessoa muito parceira, sempre muito presente, com quem eu possa confiar e conversar sobre todo e qualquer assunto; Tatiany, minha colega de laboratório e amiga querida, que aguentou toda a minha chatice e tornou o convívio dentro do laboratório muito mais leve e muito menos estressante. Vocês foram peça fundamental nesses dois anos, eu tenho um carinho muito grande por cada um de vocês. Muito obrigado, de verdade!

À Gabriela, Laryssa, Hegger e Vinícius por todas as conversas, todas as risadas, todos os choros, todos os rolês e, acima de tudo, por todo o incentivo dentro e fora do ambiente acadêmico. Contem comigo! Muito obrigado!

Aos membros do Laboratório de Bioinformática, em especial o meu orientador e amigo Dr. Glauber Wagner, pela oportunidade, paciência, ensinamentos, confiança e por ter me aceitado de braços abertos há dois anos atrás, quando eu mal sabia o que era bioinformática. Obrigado pelas nossas conversas, não apenas aquelas de cunho científico, técnico e didáticas, mas sim por todos os ensinamentos que eu vou levar comigo durante a minha vida profissional e pessoal. Obrigado por ser um orientador sempre presente, sempre preocupado, por me cobrar nos momentos necessários e por sempre me incentivar a ser um profissional melhor, da maneira mais humana e correta. Muito obrigado.

Ao "meu IC", Vilmar, por sempre estar presente quando necessário, por toda a ajuda e risadas no laboratório. Muito obrigado!

À Árvore da Vida: Guilherme, João Daniel, João Pedro, Laís e Renato. Há mais de uma década, vocês sempre estiveram presentes e, com certeza, ajudaram a construir a pessoa que eu sou hoje. Aprendi e ainda aprendo muito com todos vocês. Temos muito caminho ainda pela frente, galera. Gostaria de agradecer, em especial, toda a disponibilidade e a ajuda do Guilherme e do João Pedro neste trabalho. Muito obrigado.

Ao pessoal do Laboratório de Protozoologia: os professores Dr. Edmundo, Dr. Mario e, em especial, à professora Drª. Patrícia, pela

amizade, conselhos, puxões de orelhas, conversas, ajudas e ensinamentos ao longo de todo o meu mestrado; os alunos Abadio, Adriana, Amábili, Ana Paula, Beatriz, Bibiana, Carime, Gabriel, Iasmin, Natália, Thaís, Vanessa e Viviane. Obrigado, pessoal.

Gostaria de agradecer à Universidade Federal de Santa Catarina, por ter sido minha segunda casa nos últimos 9 anos, assim como por todo o tempo e dinheiro investido na minha formação acadêmica e pessoal. Agradecer também ao Programa de Pós-Graduação em Biotecnologia e Biociências, pela oportunidade de concluir mais uma etapa na minha caminhada de aperfeiçoamento profissional e pessoal. Agradecer o CNPq, pelo suporte financeiro e a possibilidade de ter realizado todo o meu mestrado com bolsa. Por fim, mas não menos importante, gostaria de agradecer à Alexandra Elbakyan pela iniciativa de ter criado o portal SciHub, sem o qual eu não conseguiria ter lido nem a metade dos artigos que eu li para estudar e escrever a minha dissertação, assim como por manter viva a ideia de uma ciência livre.

RESUMO

O *Trypanosoma rangeli* é um protozoário que infecta triatomíneos e diversos mamíferos para realizar o seu ciclo biológico. O *T. rangeli* tornou-se um organismo de interesse científico devido a sua similaridade genômica e proteômica com *Trypanosoma cruzi*. Do genoma de *T. rangeli* foi observado que 66% dos genes codificam "proteínas hipotéticas", que são proteínas preditas por ferramentas de bioinformática, mas que não têm suas funções caracterizadas. O estudo destes dados moleculares, através de análises computacionais comparativas, pode esclarecer os mecanismos de virulência e infectividade de outras espécies de *Trypanosoma*. Este trabalho tem como objetivo caracterizar a possível função e expressão de proteínas hipotéticas em *T. rangeli* através de análises *in silico* com base em dados genômicos, transcriptômicos e proteômicos deste organismo. Foi feita a predição de sequências a partir de diferentes dados de sequenciamento e montagens do genoma de *T. rangeli*, utilizando-se os programas Glimmer e Augustus. As 10.506 sequências proteicas preditas não redundantes foram utilizadas para realizar uma busca por similaridade com outros genomas através do algoritmo do BLAST+, com dados disponíveis no TriTrypDB v.41, das quais 6.475 encontraram correspondência de anotação, 3.740 foram anotadas como hipotéticas, 133 como pseudogenes e 158 não encontraram nenhuma correspondência, formando assim um conjunto de dados de 3.898 proteínas hipotéticas. Destas, 1.149 continham descrições ou anotações funcionais considerando os resultados do InterProScan, HMMER e RPSblast+, sendo que 788 (20,42%) destas proteínas hipotéticas continham ao menos uma descrição. Para avaliar a possível expressão destas proteínas, foram realizadas análises de evidências de expressão utilizando dados disponíveis do transcriptoma e do proteoma de *T. rangeli*. Foram encontradas 3.690 (94,66%) sequências hipotéticas com pelo menos um transcrito e 1.452 (37,25%) com pelo menos dois peptídeos nas análises de espectrometria de massas. Considerando apenas sequências que apresentavam ambas evidências de expressão, 1.018 (26,12%) sequências hipotéticas são potencialmente expressas. Finalmente, utilizando os dados gerados neste pipeline, é possível reanotar 372 (9,54%) de todas as proteínas previamente descritas como hipotéticas, pois apresentam maior respaldo para uma anotação confiável. Em conclusão, este trabalho gerou uma abordagem sistemática e integrada que permite a reanotação de proteínas *in silico* e potencialmente aplicável a outros genomas que apresentem dados de expressão.

Palavras-chave: Trypanosomatidae. Bioinformática. Genoma. Transcriptoma. Proteoma. Anotação gênica. Proteínas hipotéticas.

ABSTRACT

Trypanosoma rangeli is a protozoan that infects triatomines and mammals to complete its biological life cycle. *T. rangeli* has become an organism of scientific interest due to its genomic and proteomic similarity to *Trypanosoma cruzi*. From its genome, it was observed that 66% of the genes were annotated as "hypothetical proteins", which are proteins predicted by bioinformatics' tools, although their function is unknown. The study of this molecular data, through comparative computational analysis, may help to elucidate the mechanisms of virulence and infectivity of other Trypanosomes. This study aimed to characterize the putative function and expression of hypothetical proteins of *T. rangeli* using an *in silico* approach based on genomic, transcriptomic and proteomic data. The gene prediction was performed by Glimmer and Augustus utilizing sequenced and assembled data from different versions of *T. rangeli* genome. 10,506 non redundant protein sequences were used as query in a similarity analysis with the BLAST+ algorithm, searching against data available on TriTrypDB v. 41, of which 6,475 sequences found a hit on the database, 3,740 were annotated as hypothetical, 133 as pseudogene and 158 did not find any corresponding match, therefore forming a dataset of 3,898 hypothetical proteins. 1,149 of those had available descriptions or functional annotations considering the results found by InterProScan, HMMER, and RPSblast+, from these 788 (20.42%) hypothetical proteins had at least one description. To evaluate the possible expression of these proteins, evidence of expression analysis was performed using available transcriptome and proteomic data from *T. rangeli*. 3,690 (64.66%) protein sequences had at least one transcript associated and 1,452 (37.25%) at least two different peptides originated from a previous mass spectrometry analysis. Considering only the sequences which presented both evidence, 1,018 (26.12%) hypothetical proteins could potentially be expressed. Finally, according to the results found here, it is possible to reannotate 372 (9.54%) sequences that were previously annotated as hypothetical, as these are the sequences that show greater evidence. In conclusion, this study developed an integrated systemic analysis that allows for protein reannotation *in silico* and could be applied to other organisms that have available expression data.

Keywords: Trypanosomatidae. Bioinformatics. Genome. Transcriptome. Proteome. Genomic annotation. Hypothetical proteins.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CCB	Centro de Ciências Biológicas
CDD	Conserved Domain Database
CDS (NR)	Coding Sequence (Não redundante)
EST	Expressed Sequenced Tag
FDR	False Discovery Rate
GHMM	Generalized Hidden Markov Models
GSS	Genomic Survey Sequence
HMM	Hidden Markov Model
ITP	Integrated Transcriptomic Proteomic
MIP	Departamento de Microbiologia, Imunologia e Parasitologia
Mpb	Milhões de pares de base
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
ORF	Open Reading Frame
UniProtKB	Universal Protein Resource Knowledge Base
RNAseq	RNA sequencing
SVM	Support-Vector Machine
TPM	Transcritos Por Milhão
TriTrypDB	TriTryp Data Base
UFSC	Universidade Federal de Santa Catarina

SUMÁRIO

1 INTRODUÇÃO

1.1 A BIOINFORMÁTICA

Pode-se dizer que o nascimento da bioinformática se deu na década de 1960, quando os pesquisadores começaram a utilizar métodos computacionais para estudar dados bioquímicos sobre proteínas, uma década antes do surgimento da primeira metodologia de sequenciamento de DNA (HAGEN, 2000). Em 1977, ao mesmo tempo em que o sequenciamento de cadeias polipeptídicas era aperfeiçoado, é publicado o primeiro método de sequenciamento de DNA: o método de Maxam-Gilbert (MAXAM; GILBERT, 1977), que devido ao uso de materiais radiativos e produtos químicos perigosos não se tornou um método amplamente utilizado. Entretanto, no mesmo ano, foi descrito o método de sequenciamento de DNA através da técnica de término de cadeia, também conhecido como método de Sanger (SANGER; NICKLEN; COULSON, 1977), o qual continua a ser amplamente utilizado.

O método de Sanger revolucionou o campo da biologia computacional e da biologia molecular, possibilitando que pesquisadores investigassem a informação genética contida nos genes dos organismos. Foi assim que começaram os estudos genômicos, com os primeiros genomas completos publicados ainda na década de 1980: o DNA mitocondrial humano; do cloroplasto; o DNA da bactéria *Haemophilus influenzae*, em 1995; e o da levedura *Saccharomyces cerevisiae* (GAUTHIER et al., 2018). Ao mesmo tempo, o método de Sanger continuou sendo aprimorado para realizar o primeiro sequenciamento completo do genoma humano, porém este genoma só foi concluído e publicado em 2001 (INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM, 2001), fato este que marca o início da Era Genômica.

A partir da segunda década deste milênio, com a consolidação das tecnologias de sequenciamentos de nova geração (*New Generation Sequencing* – NGS) e a redução do custo de sequenciamento (SOUALMIA; LECROQ, 2013; SCHWARZE et al., 2018), houve uma explosão do número de sequenciamentos genéticos de diversos organismos que agora são estudados por diferentes áreas da medicina e da biologia. Atualmente, a bioinformática é o ramo das Ciências Biológicas responsável por tentar lidar com esse crescente volume de dados biológicos, a qual pode ser definida como uma área de estudo interdisciplinar que desenvolve métodos e ferramentas (programas) para análises de dados biológicos moleculares em computadores (*in silico*)

(LUSCOMBE; GREENBAUM; GERSTEIN, 2001). Basicamente, a bioinformática pode ser subdividida em três principais áreas: (i) análise de sequências biológicas, sejam elas sequências de nucleotídeos ou aminoácidos; (ii) análise de modelos tridimensionais de proteínas; e (iii) desenvolvimento de ferramentas e métodos para análise de dados biológicos (VERLI, 2014).

A maior parte das sequências biológicas são depositadas em bancos de dados públicos, como o GenBank, que é um dos principais repositórios de dados de sequências do mundo e faz parte do *National Center for Biotechnology Information* (NCBI). Atualmente, o GenBank (GenBank – NCBI, disponível em: https://www.ncbi.nlm.nih.gov/genbank) contém aproximadamente 380 bilhões de sequências nucleotídicas e mais de 660 bilhões de sequências de aminoácidos (NCBI/GenBank, acesso em 26/04/2019).

O processo pelo qual novos dados de sequências biológicas oriundas do sequenciamento de uma espécie são analisados, com a finalidade de se entender sua função e seu contexto biológico, chama-se anotação gênica (STEIN, 2001). Essa contextualização *in silico* dos dados moleculares é um dos grandes desafios para os bioinformatas que lidam com análises de sequências biológicas.

1.2 ANÁLISE DE GENOMA E ANOTAÇÃO GÊNICA

O processo de anotação de genomas pode ser subdividido em: anotação estrutural, que procura identificar as regiões funcionais e estabelecer a estrutura de exons-introns do genoma; e anotação funcional, que procura adicionar informações aos genes através da identificação de suas funções e processos biológicos do qual ele faz parte (YANDELL; ENCE, 2012). Justamente, dado um novo genoma, uma das mais importantes etapas do processo de anotação deste genoma é a determinação da estrutura e da organização dos genes que codificam proteínas (*Coding Sequences* - CDS) e caracterização das demais regiões funcionais como promotores ou terminadores (KORF, 2004). Para tal, são utilizadas ferramentas computacionais conhecidas preditores gênicos, que surgiram na década de 90 e revolucionaram a análise de genomas, possibilitando que os genes fossem encontrados de maneira rápida e fácil (YANDELL; ENCE, 2012).

Diferentes preditores gênicos utilizam algoritmos distintos para validarem as suas predições, contudo os melhores preditores gênicos são aqueles que possam ser "treinados" antes de realizarem a etapa de predição, ou seja, eles ajustam seus parâmetros para o conjunto de dados

específicos no qual será feita a predição gênica (STEIN, 2001; KORF, 2004). No início, os preditores gênicos eram conhecidos como ferramentas de análise *ab initio*, pois utilizavam apenas modelos matemáticos (como Modelos Ocultos de Markov, por exemplo) durante o seu treino (YANDELL; ENCE, 2012) e não incluíam a possibilidade de serem utilizadas evidências experimentais externas como dados de transcriptoma ou proteoma. Atualmente a maioria dos preditores gênicos utiliza uma combinação de modelos matemáticos e dados experimentais para aumentar a acurácia das predições dos genes e também da estrutura de exons-introns, quando presentes (KORF, 2004; YANDELL; ENCE, 2012).

Geralmente esses programas são a primeira etapa dentro de um pipeline científico para anotação de um genoma. Pipelines científicos são, por definição, um conjunto de elementos em séries, como processos realizados por diferentes programas, onde o resultado (*output*) de um processamento serve como entrada (*input*) do elemento seguinte (GOVERNMENT OF NEW SOUTHERN WALES, 2018). Diversos pipelines de anotação gênica foram desenvolvidos nas últimas décadas, devido a necessidade de serem anotados os genes e as estruturas de diversos novos genomas que estavam (e estão) sendo sequenciados, como: GARSA, anotação e ferramentas de análises de genomas (DAVILA et al., 2005); pipelines de anotação de pseudogenes e promotores em genomas eucariotos (SOLOVYEV et al., 2006); CEGMA, anotação de genes importantes em genomas eucariotos (PARRA; BRADNAM; KORF, 2007); KAAS, anotação de genes e reconstrução de vias metabólicas (MORIYA et al., 2007); MAKER2, anotação e ferramenta de controle de bancos de dados genômicos para NGS (HOLT; YANDELL, 2011); GENCODE, anotação de genes humanos ao genoma de referência do projeto ENCODE (HARROW et al., 2012); PROKKA, anotação de genomas de organismos procariotos (SEEMANN, 2014); STINGRAY, plataforma de análise e anotação de genomas sequenciados por Sanger ou NGS (WAGNER et al., 2014); NCBI anotação de genomas de organismos procariotos (TATUSOVA et al., 2016).

Uma vez identificado um gene, é necessário determinar sua possível função e para tal é realizada, em um primeiro momento, uma identificação desta possível função através de busca por similaridade entre as novas sequências (oriundas do sequenciamento do novo organismo) e as sequências já anotadas que estão disponíveis em bancos de dados (MORIYA et al., 2007). Essa etapa é desempenhada por algoritmos de alinhamento de sequência, como é o caso do algoritmo Smith-Waterman, utilizado pelos programas do pacote BLAST+

(BLASTn, BLASTp ou BLASTx, por exemplo) (CAMACHO et al., 2009).

Ainda, os bancos de dados utilizados neste processo devem conter dados confiáveis e, se possível, validação das anotações das sequências depositadas nestes. Neste contexto, um banco de dados de sequências bastante conhecido e utilizado pelos bioinformatas é o *Universal Protein Resource Knowledgebase* (UniProtKB, disponível em: https://www.uniprot.org/), famoso por manter um subconjunto de dados de sequências proteicas não redundantes que são revisadas ("curadas") manualmente, o Swiss-Prot. Existem também os bancos de dados especializados, sendo que para este trabalho o banco de dados mais significativo é o TriTrypDataBase (TriTrypDB, disponível em: http://tritrypdb.org/tritrypdb), que reúne sequências e informações sobre diferentes espécies de tripanossomatídeos. Tecnicamente o TriTrypDB não é considerado um banco de dados curado, porém a qualidade dos dados que são tornados públicos é revisada frequentemente uma vez que este repositório é mantido por diversos grupos de pesquisas independentes de maneira colaborativa (ASLETT et al., 2009).

1.3 HOMOLOGIA E COMPARAÇÃO DE SEQUÊNCIAS

A comparação entre sequências biológicas só é possível quando se leva em consideração o fenômeno da homologia, que presume a existência de uma relação de ancestralidade entre duas ou mais sequências ao longo da história evolutiva. Quando duas sequências biológicas compartilham uma mesma descendência, diz-se que aquelas sequências são sequências homólogas, sejam sequências nucleotídicas (genes homólogos) ou sequências aminoacídicas (proteínas homólogas) (KOONIN, 2005). Ainda, dependendo da história evolutiva dessas sequências, é possível classifica-las em ortólogas ou parálogas. Diz-se que dois genes homólogos são ortólogos quando a relação de ancestralidade entre os dois surgiu a partir de um evento de especiação, enquanto que dois genes homólogos são considerados parálogos quando estão relacionados via um processo molecular de duplicação genômica (KOONIN, 2005).

Partindo destes princípios, algoritmos que fazem comparação de sequências conseguem mensurar o grau de similaridade entre sequências ou até mesmo predizer se aquelas tais sequências são ou não são homólogas. Tipicamente, quando se comparam sequências proteicas próximas na escala evolutiva, pode-se dizer que duas proteínas têm a mesma função quando o grau de similaridade entre elas for acima de 50%,

que representa a porcentagem de resíduos de aminoácidos conservados entre as duas sequências quando comparadas (PEARSON, 2015). Esse alto grau de similaridade é o que possibilita aos bioinformatas inferirem homologias entre as sequências e realizarem o processo de transferência de anotação entre a nova sequência e outra sequência previamente anotada em banco de dados.

Entretanto, quando o grau de similaridade das sequências aminoacídicas é baixo (inferior a 30%), geralmente estes algoritmos que fazem comparação de sequência perdem sua eficiência e, então, é preciso recorrer a técnicas que busquem homologias remotas, também chamadas de homologia distantes. É preciso levar em consideração que ao longo da história evolutiva de uma determinada sequência, é possível que duas proteínas homólogas apresentem baixa similaridade, mas as estruturas e funções entre essas proteínas sejam conservadas, pois é possível que os domínios ou motivos proteicos sejam mantidos ao longo da evolução destes genes (CHEN et al., 2016).

Uma das metodologias mais comumente aplicadas para a detecção de homologias distantes é a utilização de modelos matemáticos probabilísticos, como os Modelos Ocultos de Markov (*Hidden Markov Models* – HMM), onde um algoritmo calcula uma série de modelos ou perfis de estados discretos mais prováveis e, através da comparação entre os diferentes perfis montados, procura realizar predições acerca dos padrões que possam ser encontrados dentro de um determinado conjunto de dados (KROGH et al., 1994; CHEN et al., 2016).

1.4 ANOTAÇÃO GÊNICA UTILIZANDO DADOS DE TRANSCRIPTOMA E PROTEOMA

Os avanços das tecnologias de sequenciamento de DNA foram acompanhados por avanços em outras áreas da biologia molecular, como a possibilidade de sequenciamento de alta performance de RNAs transcritos (*RNA sequencing* – RNAseq) e de se desenvolver métodos de detecção de proteínas expressas, através da espectrometria de massas, cada vez mais sensíveis. De tal forma, foi apenas uma questão de tempo para que esses dados fossem utilizados em estudos de escala genômica. Enquanto que os genes de uma espécie se mantêm relativamente imutáveis ao longo da vida de um organismo, outros processos como a formação de diferentes tipos de RNAs ou a expressão de proteínas tendem a variar significativamente para propiciar uma resposta adequada aos inúmeros estímulos e contextos ambientais os quais aquele organismo está inserido (KUMAR et al., 2016).

A maneira como os dados de transcriptoma e proteoma estão sendo utilizados para a anotação de genomas ou reanotação de genes é através de metodologias conhecidas como proteogenômicas. Tipicamente, os dados genômicos e transcriptômicos são analisados para se melhor interpretar dados proteômicos de uma determinada espécie, ao mesmo tempo em que os resultados proteômicos obtidos servem como base para a validação de dados de expressão e, também, a refinar modelos e anotação de genes (NESVIZHSKII, 2014).

Alguns pipelines já foram desenvolvidos para aplicarem os conceitos da proteogenômica, como: ITP pipeline (*Integrated transcriptomic-proteomic pipeline*), que elabora um banco de dados com informações de espectrometria de massas e realiza buscas através de montagens de transcriptomas referenciados para anotar genomas de organismos eucariotos (KUMAR et al., 2015); Galaxy-P, módulo disponível para a plataforma Galaxy Project (disponível em: https://galaxyproject.org/), que permite que análises proteogenômicas sejam realizadas *online* e de maneira mais acessível aos usuários (SHEYNKMAN et al., 2014); e Enosi, que utiliza dados de RNAseq para identificar peptídeos em dados gerados por técnicas de espectrometria de massas e, então, reanotar genomas (CASTELLANA et al., 2013).

As anotações do genoma de espécies que tenham diferentes dados biológicos disponíveis, sejam eles genômicos, transcriptômicos ou proteômicos, tendem a ganhar muito com uma abordagem "multi-ômica", principalmente as áreas do conhecimento associadas à medicina, onde a validação do conjunto desses dados pode prover um panorama mais completo aos pesquisadores e ajudar a reavaliar métodos de diagnósticos ou prognósticos de uma doença (KUMAR et al., 2016).

Mesmo com a disponibilidade de bancos de dados diversos, múltiplas ferramentas de análise e diferentes evidências experimentais, ainda existem genes que não são anotados ou são anotados de forma errônea, sem a determinação de sua função, os produtos destes genes são conhecidos como proteínas hipotéticas. Proteínas hipotéticas são aquelas cuja existência é predita através de ferramentas de predição gênica, quando da análise de um novo genoma sequenciado, mas que não têm suas funções caracterizadas. Existem ainda as proteínas hipotéticas conservadas, que são proteínas hipotéticas encontradas em linhagens filogenéticas distantes ou até mesmo não aparentadas e apresentam alto grau de similaridade entre si. Estima-se que aproximadamente 40% das proteínas preditas em um novo genoma sequenciado sejam anotadas como proteínas hipotéticas (GALPERIN, 2001). Ainda, de um ponto de vista genômico e proteômico, essas proteínas hipotéticas são bastante

importantes e relevantes para que seja possível montar uma imagem mais completa a respeito das informações moleculares de uma determinada espécie, pois elas representam possibilidades de novas funções e novas estruturas (LUBEC et al., 2005).

Diversas metodologias *in silico* podem ser aplicadas para tentar elucidar o papel das proteínas hipotéticas, porém uma das soluções propostas pela literatura é justamente o desenvolvimento de um pipeline científico que implemente diversos programas diferentes para tentar esclarecer estruturas e atribuir funções aos produtos hipotéticos, um processo mais rápido e barato do que qualquer tipo de análise experimental (FICKETT, 1996).

1.5 TRIPANOSSOMATÍDEOS

Os tripanossomatídeos são protozoários pertencentes ao Filo Euglenozoa, ordem Kinetoplastida, família Trypanosomatidae, que apresentam um único flagelo e são parasitas obrigatórios (D'ALESSANDRO, 1976). Esta família faz parte da classe Kinetoplastea, caracterizada pela presença de uma região rica em DNA extracromossomal (kDNA) chamada cinetoplasto, localizada em sua mitocôndria (D'ALESSANDRO; SARAVIA, 1999; LUKEŁ et al., 2014). Dentro deste grupo encontram-se dois gêneros de interesse biológico, pois causam patologias em humanos e animais, o gênero *Trypanosoma* e o gênero *Leishmania*. O primeiro contém espécies de parasitos que são os agentes etiológicos de doenças como a Doença de Chagas (*Trypanosoma cruzi*) na América (CHAGAS, 1909a) e a Doença do Sono (*Trypanosoma brucei*) na África (BRUCE, 1914), já o segundo grupo de parasitos são agentes etiológicos das Leishmanioses cutânea e visceral (*Leishmania* spp.) em diversas regiões do globo (LEISHMAN, 1903; STEVERDING, 2017). Entretanto, é importante destacar que a grande maioria dos gêneros de tripanossomatídeos são parasitas que acometem invertebrados (principalmente os insetos) e até mesmo plantas, como é o caso do gênero *Phytomonas* (DOLLET, 1984).

Ainda, os tripanossomatídeos apresentam diferentes formas celulares durante os seus ciclos de vida, que podem variar entre monoxeno ou heteroxeno e recebem diferentes nomenclaturas de acordo com características como: posição do cinetoplasto; a posição, inserção ou tamanho do flagelo; forma celular e de sua membrana ondulante (NEVES et al., 2011). Diversas características da família Trypanosomatidae são bem conservadas e compartilhadas entre os gêneros *Trypanosoma* e *Leishmania*, porém acredita-se que algumas

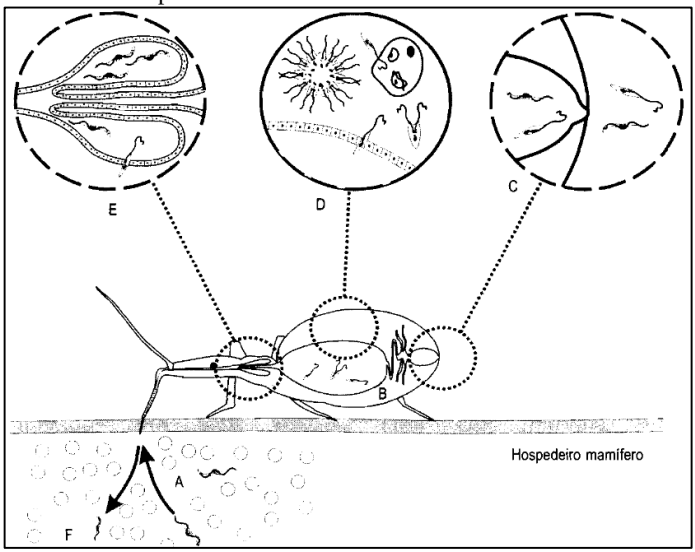
diferenças pontuais em genes específicos sejam responsáveis pela particularidade de cada parasito com relação às estratégias de sobrevivência, sua biologia e patofisiologia (EL-SAYED et al., 2005). Evidências morfológicas e moleculares apontam que os gêneros *Trypanosoma* e *Leishmania* compartilham um último ancestral comum entre 230 e 300 milhões de anos atrás, aproximadamente, com algumas divergências coincidindo com o surgimento de alguns mamíferos há 165 milhões de anos atrás. Acreditava-se que os tripanossomatídeos ancestrais utilizavam vertebrados como hospedeiros e apenas posteriormente insetos, porém evidências paleontológicas confirmam que, na verdade, os insetos foram os primeiros hospedeiros destes parasitos (LUKEŁ et al., 2014).

Dentre os membros do gênero *Trypanosoma* encontra-se o *T. rangeli*, descrito por Tejera em 1920, um organismo infectivo não-patogênico para seus hospedeiros mamíferos e que devido a sua posição taxonômica próxima à *T. cruzi*, tornou-se um organismo de interesse científico, em grande parte por conta da similaridade entre os seus materiais genéticos e constituintes antigênicos.

1.6 Trypanosoma rangeli

O *Trypanosoma (Herpetosoma) rangeli* Tejera, 1920, é um parasita que necessita de mais de um hospedeiro para completar o seu desenvolvimento, caracterizando o seu ciclo biológico como sendo heteroxeno, utilizando-se geralmente de insetos triatomíneos hematófagos dos gêneros *Rhodnius*, mas também infectando uma variedade de mamíferos de diferentes ordens (D'ALESSANDRO; SARAVIA, 1999; GUHL; VALLEJO, 2003). Resumidamente, o ciclo biológico deste parasito ocorre quando um inseto triatomíneo ingere as formas tripomastigotas do parasito presentes na corrente sanguínea de um mamífero infectado. Estes protozoários se desenvolvem no intestino do inseto, em sua forma epimastigota, as quais penetram as paredes do intestino, passam à hemocele e alcançam a hemolinfa, migrando em direção às glândulas salivares onde se diferenciam em sua forma tripomastigotas metaciclíca. Então, estas formas são transmitidas ao novo hospedeiro mamífero durante o processo de repasto de sangue do inseto (Figura 1).

Figura 1. Desenho esquemático do ciclo de vida do *T. rangeli* no hospedeiro invertebrado e no hospedeiro mamífero.



A) Ingestão das formas tripomastigotas de *T. rangeli* durante o repasto sanguíneo do triatomíneo num hospedeiro infectado; B) Formas tripomastigotas e epimastigotas no trato intestinal do triatomíneo; C) Parasitos alcançam a ampola retal; D) Adesão de formas epimastigotas do parasito ao epitélio intestinal e transposição para a hemocele, onde se multiplicam; E) Invasão das glândulas salivares do triatomíneo pelo parasito e diferenciação para formas infectantes (tripomastigotas metacíclicos); F) inoculação das formas infectantes através da saliva durante o repasto sanguíneo.
Fonte: GRISARD; STEINDEL, 2004 (modificado).

1.7 GENOMA, TRANSCRIPTOMA E PROTEOMA DE *Trypanosoma rangeli*

Apesar de não ser patogênico para seres humanos, a similaridade antigênica, genética e a distribuição simpátrica entre *T. cruzi* e *T. rangeli* confere problemas nos diagnósticos da Doença de Chagas (AFCHAIN et al., 1979; MORAES et al., 2008). Com isso, estudos de genômica, proteômica e transcriptômica de *T. rangeli* vêm sendo desenvolvidos no intuito de encontrar novos marcadores para diagnóstico, assim como para obter um melhor entendimento de sua biologia, genética e diversidade. Este contexto culminou na publicação da primeira versão do genoma de *T. rangeli* por STOCO e colaboradores (2014), que começou quando SNOEIJER e colaboradores (2004) publicaram um estudo piloto

disponibilizando os primeiros Marcadores de Sequência Expressa (*Expressed Sequence Tag* – EST). Depois, entre 2006 e 2013, uma série de estudos genômicos e proteômicos publicaram uma série de ESTs e outras Buscas de Sequências Genômicas (*Genome Survey Sequence* – GSS) (WAGNER, 2006; FERREIRA et al., 2010; GRISARD et al., 2010), assim como dados de espectrometria de massas total e de superfície (LUCKEMEYER, 2006; WAGNER et al., 2013).

As informações disponíveis a respeito do genoma de *T. rangeli* SC58 demonstram que este parasito tem um genoma com tamanho de aproximadamente 24 milhões de pares de bases (Mpb), com um conteúdo GC de 49,91% (STOCO et al., 2014). O sequenciamento deste organismo apresenta uma cobertura genômica de 13,8x e foi feito utilizando-se a plataforma de sequenciamento Roche-454, sendo que a montagem desses dados gerou aproximadamente 259 scaffolds que apresentavam um N50 de 202.734 pares de bases (pb) (STOCO et al., 2014). No caso do *T. rangeli* foram encontradas 7.613 CDS em seu genoma, que apresentam um conteúdo GC de 54,27% (STOCO et al., 2014). É importante destacar que a organização dos genes nos parasitos do gênero *Trypanosoma* e *Leishmania* relembra uma organização gênica tipicamente bacteriana, que quando as unidades funcionais do genoma são transcritas produzem uma série de pré-mRNAs de maneira policistrônica (TEIXEIRA et al., 2012). Notavelmente, *T. rangeli* compartilha até 93% de seus genes com outros *Trypanosoma* patogênicos, sendo que 2.414 (31,71%) de todas as CDS codificam proteínas diversas e 5.043 (66,24%) são proteínas anotadas como "proteínas hipotéticas" (STOCO et al., 2014).

1.8 JUSTIFICATIVA

Na grande maioria dos estudos genômicos as proteínas anotadas como hipotéticas não são revisadas e são desconsideradas das análises, justamente pela sua natureza incerta, portanto incluir essas proteínas nos estudos tende apenas a aumentar a riqueza e a autenticidade dos trabalhos. Desta forma, um pipeline científico que possa ser utilizado para reanotar proteínas hipotéticas em diversos genomas diferentes pode abrir várias possibilidades de novas análises e estudos sobre a função desconhecida de muitos produtos gênicos.

1.9 HIPÓTESE

Abordagens ômicas integradas diminuem o número de proteínas anotadas como hipotéticas em um genoma.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Desenvolver uma ferramenta para atribuir anotações às proteínas hipotéticas através de análises *in silico* com base em dados genômicos, transcriptômicos e proteômicos.

2.2 OBJETIVOS ESPECÍFICOS

- Analisar os diferentes modelos de predição gênica baseando-se em dados genômicos disponíveis;
- Realizar a anotação funcional das proteínas hipotéticas com base na anotação funcional relacional InterProScan: IPR e Gene Ontology;
- Buscar homologias distantes das proteínas hipotéticas, através do uso de modelos matemáticos probabilísticos ou de matriz de escores que auxiliem no encontro de possíveis homólogos e domínios de famílias proteicas conservadas, assim como seus diferentes motivos;
- Avaliar a expressão dos produtos gênicos hipotéticos utilizando dados de transcriptoma e proteoma;
- Desenvolver um pipeline automatizado para anotação de proteínas hipotéticas que possa ser aplicado para diferentes organismos.

3 MATERIAL E MÉTODOS

As análises desenvolvidas neste trabalho foram realizadas no Laboratório de Bioinformática do Departamento de Microbiologia, Imunologia e Parasitologia (MIP), do Centro de Ciências Biológicas (CCB) da Universidade Federal de Santa Catarina (UFSC). Nossa estrutura conta também com servidores alocados na Superintendência de Governança Eletrônica e Tecnologia da Informação e Comunicação (SeTIC) da UFSC para desenvolvimento de plataformas computacionais e demais análises de alta performance.

3.1 DESENHO EXPERIMENTAL

Uma visão geral de todos os processos e programas utilizados neste trabalho estão incluídos e esquematizados como etapas dentro de um pipeline científico, conforme mostra a Figura 2.

Os scripts que compõe as etapas de preparação, tratamento e análise dos dados implementados neste pipeline foram escritos nas linguagens de programação Perl ou Python 3. Uma listagem de todos os programas utilizados neste pipeline encontra-se na Tabela 1.

Figura 2. Desenho experimental do pipeline científico desenvolvido neste trabalho.

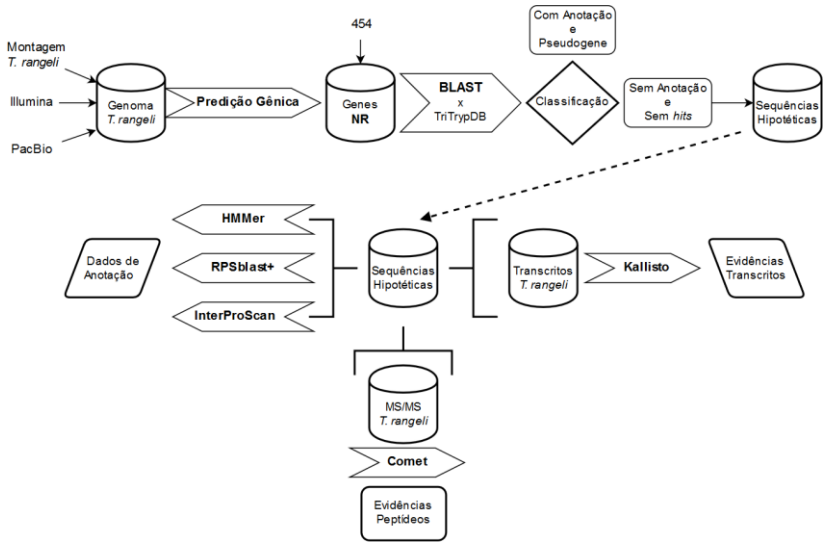


Tabela 1. Listagem de programas implementados no pipeline de anotação de proteínas hipotéticas.

Etapa	Programa	Referência
Predição Gênica	Glimmer	DELCHER et al., 2007
	Augustus	STANKE; MORGENSTERN, 2005
Agrupamento de sequências	CD-Hit	LI; GODZIK, 2006
Análise de Similaridade	BLAST+	CAMACHO et al., 2009
Análise Integrada de Anotação e Função	InterProScan	JONES et al., 2014
	HMMER	EDDY, 2015
	RPSblast+	CAMACHO et al., 2009
Mapeamento de Transcritos e Peptídeos	Kallisto	BRAY et al., 2016
	Comet	ENG; JAHAN; HOOPMANN, 2012

3.2 DADOS MOLECULARES DE *Trypanosoma rangeli*

Os dados genômicos utilizados neste trabalho foram obtidos a partir de três diferentes métodos de sequenciamento, sendo eles: sequências gênicas oriundas do sequenciamento através da plataforma 454 (STOCO et al., 2014); genoma montado a partir de sequenciamento utilizando-se a plataforma Illumina; e genomas montados que foram gerados na plataforma PacBio. É importante destacar que os dados genômicos de *T. rangeli* SC58 nas plataformas Illumina e PacBio foram produzidos em colaboração com o Karolinska Institutet, na Suécia, e gentilmente cedidos pelo Prof. Dr. Edmundo Carlos Grisard, da Universidade Federal de Santa Catarina.

Previamente às análises deste trabalho, todas essas diferentes versões do genoma do parasito foram montadas, de forma que as sequências obtidas a partir da plataforma 454 foram utilizadas já na forma de CDS, conforme descrito no genoma publicado por Stoco e colaboradores (2014), as sequências do Illumina na forma de contigs e as sequências obtidas utilizando PacBio na forma de scaffolds, bem como uma montagem do genoma em que foram utilizadas diferentes abordagens de sequenciamento, gentilmente disponibilizada por

colaboradores da Universidade Federal de Minas Gerais (dados não publicados). Para esta montagem mais recente foi dado o nome de "Montagem *T. rangeli*" e todas as principais características e estatísticas de todas as montagens utilizadas neste trabalho estão descritas na Tabela 2.

Tabela 2. Tabela com os dados sobre os diferentes sequenciamentos e montagens do genoma de *T. rangeli* utilizados neste trabalho.

	Genoma *T. rangeli* *	**Montagem Illumina**	**Montagem PacBio**	**Montagem *T. rangeli***
Conteúdo GC	55 %	52,4 %	53,6 %	53,5 %
Tamanho máximo	13.347	1.144.937	12.576.647	12.509.784
Tamanho médio	1.377	79.631	213.855	212.868
Tamanho mínimo	153	1.601	14.135	14.133
Bases A	22,12 %	23,00 %	23,12 %	23,16 %
Bases C	25,22 %	25,42 %	26,73 %	26,69 %
Bases G	29,76 %	24,69 %	26,75 %	26,72 %
Bases T	22,90 %	22,46 %	23,22 %	23,26 %
Bases N	< 1 %	4,42 %	0,18 %	0,18 %
Número de Bases	10.274.856	20.624.678	25.662.659	25.544.173
Número de Sequências	7.457 CDS	259 Contigs	120 Scaffolds	120 Scaffolds

* Stoco et al., 2014

Além dos dados genômicos, foram utilizados dados de transcriptoma e de proteoma de *T. rangeli*. Os dados de transcriptoma de *T. rangeli* utilizados neste trabalho foram gerados pelo grupo de pesquisa do Laboratório de Protozoologia da UFSC em parceria com o Prof. Dr. Guilherme Toledo e foram gentilmente cedidos pela Profª. Drª. Patrícia Hermes Stoco. Rapidamente, a partir de formas epimastigotas e trypomastigotas de *T. rangeli* cepa Choachí, coletadas de diferentes amostras biológicas, foram sequenciadas as bibliotecas de cDNA através da plataforma Illumina. A montagem deste transcriptoma foi feita anteriormente a este trabalho, utilizando-se o programa Trinity, na versão 2.2.0 (GRABHERR et al., 2011).

Os dados de proteoma utilizados neste trabalho representam o proteoma de superfície e o proteoma total do parasito, os quais foram gerados a partir de análises de espectrometria de massas com amostras de cultura de *T. rangeli* Choachí nas formas epimastigotas e trypomastigotas, em diferentes meios e com diferentes preparações, conforme descritos nos trabalhos de Wagner (2006), Wagner e colaboradores (2013) e Lückemeyer (2014).

3.3 PREDIÇÃO GÊNICA

Foram realizadas predições gênicas em todos os conjuntos de dados, exceto nas CDS obtidas do sequenciamento com o Roche 454, utilizando duas ferramentas: Glimmer (versão 3.02b) e Augustus (versão 3.2.3).

Para treinar um modelo matemático de Cadeias de Markov que seria utilizado pelo Glimmer, através do algoritmo `build-icm` com parâmetros padrões, foram utilizadas 70.017 sequências nucleotídicas (CDS) de sete genomas de tripanossomatídeos disponíveis no TriTrypDB v.41 (*T. brucei gambiense*; *T. brucei* Lister; *T. brucei* TREU927; *T. cruzi* CL Brener Esmeraldo-like; *T. cruzi* CL Brener Non-esmeraldo-like; *T. cruzi marinkellei*; *T. grayi*). Em seguida, este modelo treinado foi aplicado para a predição dos CDS de *T. rangeli* pelo programa, considerando que o genoma deste organismo é um genoma linear (`--linear`), onde não há sobreposição de genes (`--max_olap` 0), dando preferência para os CDS que iniciem com códon ATG (`--start_codon` atg) e desconsiderando genes preditos menores do que 300 pares de bases (`--gen_len` 300).

Já para o programa Augustus, a fim de gerar um modelo de treino necessário para a predição dos genes, foi utilizado como referência a montagem do genoma de *T. rangeli* oriunda do sequenciamento por PacBio e como evidência de expressão foram utilizadas as proteínas descritas no genoma publicado por Stoco e colaborades (2014) e aquelas identificadas por espectrometria de massas descritas no trabalho publicado por Wagner e colaboradores (2013). O conjunto de dados de treino foi obtido através da plataforma online Web Augustus (disponível em: http://bioinf.uni-greifswald.de/webaugustus/). Em seguida, este modelo criado foi utilizado para realizar a predição dos genes, considerando apenas a característica do genoma linear de *T. rangeli* como parâmetro neste sistema (`--genemodel` = intronless).

Por fim, as sequências dos genes preditos foram avaliadas quanto a quantidade de bases sem identificação (N), sendo removidas aquelas sequências que continham mais do que 10% do seu conteúdo total como bases N. Este tratamento foi feito por um script *in-house*.

3.4 AGRUPAMENTO DOS GENES PREDITOS

Com o objetivo de diminuir o número de sequências redundantes identificadas pelos preditores gênicos, foi realizada uma etapa de agrupamento das sequências obtidas. O programa escolhido para tal foi o CD-Hit, pois ele agrupa as sequências em grupos (*clusters*) não redundantes a partir de alinhamentos local par-a-par de cada sequência, gerando um conjunto de dados não redundantes, ao qual atribuímos o nome de "CDS-NR". Para tal, foram agrupadas sequências que continham similaridade maior do que 95% em *clusters*, levando em consideração o maior tamanho-de-palavra (*word size*) recomendado para tal valor limítrofe. Este *word size* é variável entre sequências nucleotídicas e sequências aminoacídicas, sendo que no presente trabalho foram utilizados os *word size* "10" e "5", respectivamente.

O objetivo principal da etapa de agrupamento é nos certificarmos de que estávamos trabalhando apenas com sequências únicas, para diminuirmos as chances de obtermos resultados repetitivos no restante das análises do pipeline. A partir deste ponto todas as demais etapas foram realizadas com sequências de aminoácidos preditas não redundantes, exceto pelas análises dos dados de transcriptoma que utilizaram as sequências de nucleotídeos preditas não redundantes.

3.5 ANÁLISES DE SIMILARIDADE

De posse do conjunto de dados de sequências não-redundantes, foi realizada uma etapa de análise de similaridade para buscar possíveis anotações depositadas em bancos de dados para essas sequências, com o intuito de realizar um processo de transferência de anotação através da similaridade entre as sequências. O programa utilizado para as análises de similaridade de sequências foi o BLAST+, através do algoritmo BLASTp (CAMACHO *et al*., 2009).

O banco de dados escolhido para esta etapa do trabalho foi o TriTryDB, sendo que foram apenas utilizados os dados de sequências públicas que não estão protegidos por direitos autorais, totalizando 28 genomas de tripanossomatídeos analisados que correspondem a 280.312 sequências aminoacídicas. As sequências utilizadas neste trabalho

correspondem à versão 41 do TriTrypDB e foram adquiridas no dia 04 de janeiro de 2019.
A etapa de análise de similaridade foi realizada utilizando-se um valor de *e* de `1e-5` e número máximo de resultados (*hits*) por sequência (*query*) igual à `15`. Os resultados obtidos foram classificados em quatro categorias: 1) sequências "com anotação"; 2) sequências cuja anotação era "pseudogene"; 3) sequências hipotéticas, cuja anotação continha uma ou mais palavras-chave (*fragment*, *hypothetical* ou *partial*); e 4) sequências sem resultados de similaridade no banco de dados.

As sequências que foram classificadas na primeira categoria (sequências com anotação) tiveram seus parâmetros de `identidade` e `positividade` reavaliadas, através de um script *in-house*, que utilizou valores limítrofes (`37,54` e `54,48` respectivamente) para reclassificá-las como sequências pertencentes à categoria de sequências hipotéticas.

3.6 ANOTAÇÃO INTEGRADA DAS SEQUÊNCIAS HIPOTÉTICAS

A primeira análise que diz respeito às possíveis anotações e às possíveis funções das sequências de proteínas hipotéticas foi feita utilizando o programa InterProScan (versão 5.33-72.0). Para este trabalho, o InterProScan foi executado utilizando os parâmetros padrões. Em função de não nos permitir ajustar nenhum parâmetro estatístico para avaliar a quão restritiva ou flexiva foi realizada a análise, outras duas análises com relação a possíveis anotações e funções dos produtos gênicos foram realizadas de maneira independente. É importante destacar que os bancos de dados associados com informações estruturais das proteínas (como MobiDBLite e Coils) foram removidos das análises, pois não contém informações a respeito da anotação ou função das sequências de interesse.

3.6.1 Busca por homologias distantes

Com o conjunto de sequências de proteínas hipotéticas foram realizadas análises que aplicam o método da busca de homologias distantes através de modelos matemáticos probabilísticos (perfis de *HMM*) associados às sequências através de algoritmos disponíveis no pacote do HMMER (versão 3.1b2). Para realizar estas análises, foi utilizado o algoritmo `hmmscan` do HMMER, considerando os valores de *e* que foram utilizados tanto para as sequências quanto para os domínios encontrados pelo programa, com os seguintes argumentos,

respectivamente: `-E 1e-5` e `--domE 1e-5`, contra o banco de dados do Pfam (versão 32.0 lançada em setembro de 2018, disponível em: https://pfam.xfam.org/).

3.6.2 Investigação de domínios conservados

A última parte destas análises, que dizem respeito às possíveis anotações e funções dos produtos gênicos hipotéticos, foi um estudo dos domínios conservados que estão presentes nessas proteínas preditas. O programa utilizado nesta etapa foi o RPSblast+, o qual faz parte do pacote do BLAST+, sendo que na sua execução foram utilizados os parâmetros padrões, a não ser pelo valor de *e* que foi alterado para `1e-5`, contra o banco de dados do *Conserved Domain Database* (CDD, versão publicada em março de 2014, disponível em: https://www.ncbi.nlm.nih.gov/cdd/).

3.7 QUANTIFICAÇÃO DE TRANSCRITOS

A etapa de análise que diz respeito às evidências de transcrição do conjunto de produtos gênicos hipotéticos foi realizada através dos algoritmos do programa Kallisto. É importante deixar claro que esta etapa foi a única análise de todo o pipeline científico que utilizou sequências nucleotídicas.

Através do algoritmo `kallisto-index`, foi criado um índice de todas as sequências hipotéticas utilizando os parâmetros padrões. Depois, este índice de sequências foi utilizado como base para que os *reads* (tanto *forward* quanto *reverse*) oriundos dos resultados de RNAseq fossem alinhados e quantificados através do algoritmo `kallisto-quant`, utilizando-se parâmetros padrões a não ser o número de amostragens (*bootstraps*), que foi de `5000`. Dessa forma as proteínas hipotéticas foram classificadas de acordo com a presença ou ausência de transcritos alinhados, os quais eram contabilizados na forma de transcritos por milhão (TPM).

3.8 AVALIAÇÃO DA DETECÇÃO DE PEPTÍDEOS

O último processo realizado pelo pipeline diz respeito a validação dos produtos gênicos hipotéticos através do programa Comet que, basicamente, compara as sequências proteicas contra um conjunto de dados experimentais de peptídeos obtidos por análises de espectrometria de massas.

A execução do programa foi realizada modificando-se os seguintes parâmetros no arquivo de configurações do programa (`comet.params`): pesquisa de peptídeos em bancos de dados de proteínas invertidas (`decoy_search = 1`); "Daltons" como unidade de massa molecular (`peptide_mass_units = 1`); variação aceitável da massa molecular do peptídeo (`peptide_mass_tolerance = 0.50`); busca por cisteínas carbometiladas, em função da preparação das amostras por redução e alquilação (`variable_mod02 = 57.02146 C 0 4 -1 0 0`); janela de análise experimental (`scan_range = 200 2000`).

Ainda, os resultados obtidos foram reavaliados por um script *in-house* que considerou como resultados positivos para evidências de tradução apenas as sequências que: (i) apresentavam 2 ou mais peptídeos não redundantes; (ii) não apresentassem correspondência com o banco de dados de proteínas invertidas (resultados *decoys*); (iii) tivessem carga iônica acima de 1; e (iv) parâmetro de `xcorr` maior do que `0.4`. Por fim, as sequências foram classificadas de acordo com a presença ou ausência de correspondência com os peptídeos comparados.

3.9 VALIDAÇÃO ESTATÍSTICA DOS DADOS

A análise estatística dos dados já está incluída dentro dos próprios programas utilizados neste pipeline, sendo que no caso específico das análises de parâmetros dos programas do pacote BLAST+ foi utilizado o software Statistica 6.0® como uma primeira análise e, num segundo momento, duas bibliotecas científicas de Python3: SciPy, para realizar o teste estatístico de D'Agostino-Pearson e confirmar os resultados preliminares encontrados pelo Statistica6.0®; e Matplotlib para criação dos gráficos.

3.10 PLATAFORMA COMPUTACIONAL

Os experimentos computacionais foram realizados em Servidores Dell, contendo 40 núcleos de processadores (3.2 GHz), com 320 GB de RAM (DDR4, 2400 MHz) e armazenamento de 5 TB (HDD SATA 2.5" 7200 RPM).

4 RESULTADOS E DISCUSSÃO

4.1 PIPELINE PARA REANOTAÇÃO DE GENES HIPOTÉTICOS

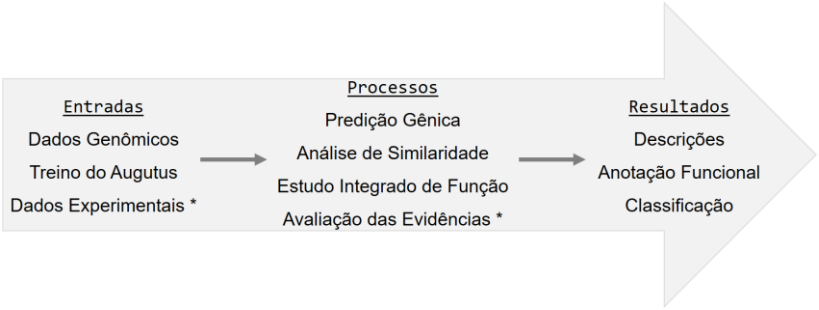
O processo de anotação ou reanotação dos genes e de seus produtos é uma tarefa que requer uma grande quantidade de análises, geralmente atreladas ao uso de diversos programas e bancos de dados disponibilizados pela comunidade científica, que permitem ao pesquisador tanto em uma via de rede mundial de computadores (*online*) ou localmente em sua própria máquina. Evidente que estas análises são simples de serem realizadas quando há um ou poucos produtos para serem estudados, porém quando existe a necessidade de analisar um genoma ou um proteoma inteiro, esta abordagem torna-se inviável.

Para contornar esse problema, surgem os *workflows*, plataformas ou pipelines científicos (ROMANO; MARRA; MILANESI, 2005), que são sistemas com vários programas que realizam as análises em um conjunto de proteínas de forma ordenada e contínua, com mínima intervenção dos usuários, gerando resultados compilados e, em muitos casos, disponibilizando informações para o usuário de maneira automatizada. Embora esta solução pareça ser simples, um outro problema enfrentado pelos usuários é a grande diversidade de formatos de dados gerados durante uma análise (RICE; LONGDEN; BLEASBY, 2000), muitas vezes necessitando adaptações para que a saída de um programa possa ser utilizada como entrada em outro programa.

Apesar de existirem um grande número de pipelines com a finalidade de realizar anotação de genomas (DAVILA et al., 2005; SOLOVYEV et al., 2006; PARRA; BRADNAM; KORF, 2007; MORIYA et al., 2007; HOLT; YANDELL, 2011; HARROW et al., 2012; SEEMANN, 2014; WAGNER et al., 2014; TATUSOVA et al., 2016), pipelines que desempenhem tarefas específicas são importantes para a comunidade científica e é neste contexto que surge a proposta de desenvolvimento deste trabalho: desenvolver um processo que possibilite a reanotação de proteínas hipotéticas de diferentes genomas, com base em evidências de expressão.

No presente pipeline, foram integrados nove programas (Tabela 1) através de 18 scripts para o tratamento de dados de entrada (*input*), execução dos diferentes programas, tratamento dos resultados obtidos e geração dos resultados (*output*) (Figura 3). Este pipeline é executado em ambiente *shell* (Unix).

Figura 3. Representação gráfica do procedimento geral deste pipeline.



Estão demonstrados todos os arquivos de entrada, os processos realizados e os resultados finais, incluindo os arquivos e análises opcionais (marcadas com um asterisco).

4.1.1 Input

Na entrada do pipeline é necessário a submissão de dados de DNA genômicos em formato `FASTA`. Como dados de entrada, sequências completas e sem espaços (*gaps*) são ideais, mas o *input* típico seria um conjunto de sequências contíguas longas oriundas de uma ferramenta de montagem de genoma. Caso o usuário deseje utilizar o programa Augustus para algum organismo cujo programa não está previamente treinando é necessário fornecer os parâmetros de teste do mesmo ou, caso deseje utilizar um modelo já incluído no Augustus, basta informar o nome do modelo a ser utilizado. Todos esses arquivos de entrada são obrigatórios para a execução do pipeline.

4.1.2 Análises opcionais

Opcionalmente, como parte das análises realizadas pelo pipeline, é possível que o usuário forneça dados de transcriptoma (*reads* limpos em formato `FASTQ`) e proteoma (espectros de massa MS/MS em formato `mzXML`) do organismo. Apesar desta etapa não ser obrigatória, dados da literatura evidenciam que a identificação e classificação das proteínas é aprimorada quando utilizadas evidências de expressão (CASTELLANA et al., 2013; NESVIZHSKII, 2014), desta forma recomenda-se a inclusão destes dados de entrada. Neste caso, as análises de transcriptoma serão realizadas pelo programa Kallisto, que quantificará a quantidade de transcritos para cada sequência e fornecerá um valor em transcritos por milhão (TPM) para uma determinada sequência. As análises de proteoma

serão realizadas pelo programa Comet, que utilizará os espectros de massa (MS/MS) para a identificação *de novo* de peptídeos e usá-los para mapear estes peptídeos nas sequências proteicas de interesse.

4.1.3 Output

Ao final da execução, dentro da pasta "Output", o usuário encontrará um arquivo em formato TSV que contém todas as sequências, suas possíveis descrições e suas possíveis anotações funcionais. Caso o usuário opte por fazer as análises utilizando evidências experimentais neste resultado constará também a possível quantificação de transcrição das sequências e de peptídeos identificados. Este resultado pode ser facilmente consultado através de um programa de editor de planilhas, como o Microsoft Excel®.

Ainda, são disponibilizadas quatro classificações das sequências contidos em arquivos de formato FASTA: (1) sequências com anotação e com evidência experimental; (2) sequências com anotação e sem evidência experimental; (3) sequências sem anotação e com evidência experimental; e (4) sequências sem anotação e sem evidência experimental.

4.1.4 Considerações

Neste contexto, é possível dizer que este pipeline será mantido e aperfeiçoado pelo Laboratório de Bioinformática, com novos algoritmos, programas e scripts, à medida em que a pesquisa em anotação automática continua.

4.2 REANOTAÇÃO DE PROTÉINAS HIPOTÉTICAS DE *Trypanosoma rangeli*

Para o desenvolvimento do pipeline, utilizamos como modelo o genoma de *T. rangeli* cepa SC58. Este foi escolhido pela disponibilidade de diferentes versões do seu genoma, de dados de transcriptoma e de proteoma. Assim, permitindo a comparação dos resultados de cada etapa do pipeline com o genoma publicado (STOCO et al., 2014). Além disto, a grande quantidade de genes anotados como hipotéticos no genoma deste parasito faz com que este seja o modelo perfeito para avaliar a funcionalidade e a proposta deste pipeline. Desta forma, iremos descrever e discutir cada etapa, parâmetro e resultados do pipeline com base nos dados deste modelo biológico.

4.2.1 Predição de genes não redundantes

Todo o conhecimento genômico de um organismo começa pelo estudo da organização dos genes, seja buscando entender a maneira como eles funcionam ou pela maneira como estão estruturados. Esses dados possibilitam que sejam elucidadas algumas das principais funções biológicas desempenhadas por um organismo, sendo que o processo de predição dos genes é a etapa mais evidente durante o processo de anotação genômica de um novo genoma (STEIN, 2001). Nos últimos anos é perceptível o desenvolvimento de novas ferramentas de predição gênica, principalmente por causa do aumento considerável na quantidade de projetos de sequenciamento genômicos que estão sendo realizados e do crescente número de sequências completas depositadas em bancos de dados. Realizar a anotação de genomas é uma tarefa bastante desafiadora, a qual só será possível utilizando-se o auxílio de ferramentas computacionais (PARRA; BRADNAM; KORF, 2007), portanto a escolha do preditor gênico correto é fundamental para se obter maior acurácia dos genes preditos.

Considerando a importância do processo de predição gênica, decidimos avaliar duas ferramentas de predição que são utilizados em diversos trabalhos genômicos: Glimmer e Augustus (D'ARGENIO; SALVATORE, 2015; SUN et al., 2015; GALLONE et al., 2016; GAO et al., 2019). Ambos se apresentam com propostas distintas: o Glimmer é um preditor gênico *ab initio* voltado para a predição gênica em organismos procariotos, que utiliza modelos matemáticos para criação de um conjunto de treino antes da aplicação deste para a predição gênica, sendo que esses modelos consideram apenas as características intrínsecas dos genes (como tamanho, conteúdo GC, tipos de códons, etc); enquanto o Augustus é um preditor gênico que, apesar de também ser um programa que se utiliza de modelos matemáticos para a predição *ab initio*, considera evidências experimentais para a criação dos parâmetros de treino, podendo ser considerado um dos principais preditores gênicos utilizado para análises em genomas de organismos eucariotos.

Desta forma, foram realizadas predições gênicas em todas as montagens do genoma de *T. rangeli* disponíveis, com exceção dos dados oriundos do sequenciamento do 454 que já são ORFs (*Open Reading Frames* – Janelas Aberta de Leitura) preditas e publicadas no genoma deste parasito por Stoco e colaboradores (2014). A quantidade de genes preditos pelo Glimmer e pelo Augustus foi de: 9.827 e 7.931, na montagem Illumina; 12.303 e 10.483 na montagem PacBio; e 11.768 e

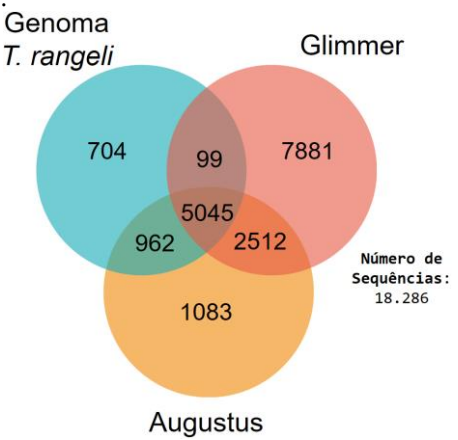
10.039 na montagem mais recente, a qual foi criada utilizando dados da montagem PacBio e da montagem Illumina. Totalizando 33.898 genes preditos pelo Glimmer e 28.453 genes preditos pelo Augutus.

Em seguida todas essas sequências preditas passaram por um controle de qualidade com o intuito de remover sequências que continham um número de bases indeterminadas (N) acima de 10% em relação ao seu tamanho. Este tratamento foi realizado por um script *in-house* e resultou em uma maior perda para aqueles genes que foram preditos a partir dos dados da montagem do sequenciamento Illumina. Isto pode ser explicado devido a presença de artefatos de sequenciamento inerentes da técnica de sequenciamento da plataforma Illumina, pois processos químicos imperfeitos envolvidos nesse processo, tanto no momento do sequenciamento por síntese quanto pela leitura do produto do sequenciamento, acabam por aumentar os erros de *base calling* nas porções finais dos *reads* (LEDERGERBER; DESSIMOZ, 2011).

De posse de 62.351 genes preditos e tratados, foi realizada uma etapa de agrupamento para obter um conjunto de genes não redundantes. Para determinar o melhor parâmetro de identidade que seria utilizado para agrupar as nossas sequências, diversos testes foram realizados com valores de identidade distintos (90%, 95%, 99% e 100% de identidade) em virtude da diversidade dos dados genômicos que foram utilizados. Em virtude das diferenças genômicas existentes entre os organismos, não há um consenso na literatura sobre qual valor utilizar para realizar esta etapa de agrupamento (KANEHISA; SATO; MORISHIMA, 2016; PAEZ-ESPINO et al., 2016; PEDERSEN et al., 2016). Após esses testes, foi determinado que o valor de 95% de identidade apresentou os melhores resultados de agrupamento para as nossas sequências. Desta forma, foram geradas um total de 15.550 sequências preditas não redundantes pelo programa Glimmer e 9.694 sequências preditas não redundantes pelo programa Augustus.

Em seguida, uma nova etapa de agrupamento foi realizada, desta vez considerando os genes preditos pelo Glimmer, os genes preditos pelo Augustus e as ORFs descritas no genoma de *T. rangeli* (STOCO et al., 2014). A Figura 4 apresenta o número de sequências não redundantes formadas pela combinação das sequências das três diferentes origens, bem como o número total de sequências não redundantes, que foi de 18.286 sequências.

Figura 4. Diagrama de Venn com o número total de sequências oriundas das três diferentes origens.



Subgrupos: oriundas do genoma de *T. rangeli* na plataforma 454 (azul claro), do preditor gênico Glimmer (vermelho claro) e do preditor gênico Augustus (amarelo), assim como o total de sequências.

Diante do resultado obtido nesta etapa, decidiu-se pela exclusão dos dados do Glimmer, devido: (1) a grande quantidade de sequências (43,10% do total de sequências não redundantes) preditas exclusivamente por este programa; (2) por estas sequências únicas apresentarem tamanho inferior a 300 pares de bases; e (3) ausência de similaridade de grande parte dessas sequências contra o banco de dados de tripanossomatídeos.

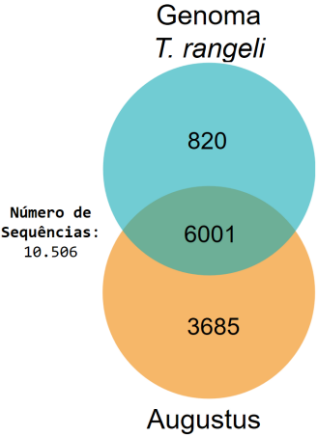
O Glimmer é um programa descrito como preditor gênico *ab initio* para genes de organismos procariotos (DELCHER et al., 2007). Este programa leva em consideração uma série de características intrínsecas dos genes, a partir de um conjunto de genes utilizados para o treinamento, como seu tamanho e o conteúdo GC. A escolha do Glimmer como um dos programas de predição gênica deste trabalho se justificava pelo fato de que a estrutura gênica em *T. rangeli* ser similar àquela encontrada em organismos procariotos. A transcrição das unidades funcionais do genoma deste parasito não é regulada no início do processo, mas sim em um nível pós-transcricional, além do fato de que os genes que codificam para proteínas sejam transcritos de maneira policistrônica (BEN-DOV; LEVIN; VÁZQUEZ, 2005) e, também, não há presença de íntrons no genoma deste tripanossomatídeo.

Por outro lado, o programa Augustus é um preditor gênico comumente utilizado para predição de genes em organismos eucariotos. Este faz uso de modelos matemáticos probabilísticos generalistas

associados aos Modelos Ocultos de Markov (*Generalized Hidden Markov Models* – GHMM) para definir diversas características em uma determinada sequência gênica: exons, introns, regiões intergênicas, regiões UTR e outros (STANKE; MORGENSTERN, 2005). A principal diferença e vantagem do Augustus com relação ao Glimmer é justamente o fato de que o Augustus pode receber evidências de transcrição, como ESTs ou de tradução, como proteínas previamente descritas daquele determinado organismo.

A partir desta decisão apenas as sequências não-redundantes preditas pelo Augustus e os dados do sequenciamento pela plataforma 454 seriam utilizadas no restante das análises deste pipeline, em um total de 10.506 sequências que passamos a descrever como CDS-NR (Figura 5). O número de CDS-NR está de acordo com o que há descrito na literatura para este parasito (STOCO et al., 2014) e outras espécies filogeneticamente aparentadas (EL-SAYED et al., 2005). Nota-se também que o número de sequências um pouco acima dos números descritos pela literatura citada se deve, principalmente, ao fato de que foram utilizadas montagens diversas em um mesmo processo de predição de sequências. Na Tabela 3 é possível observar a similaridade entre os dados oriundos do sequenciamento 454 e das CDS-NR obtidas neste trabalho, tanto em relação ao tamanho médio das sequências, como seu conteúdo GC e frequência de bases nitrogenadas em cada conjunto de sequências.

Figura 5. Diagrama de Venn do número total de sequências preditas não redundantes.



Subgrupos: oriundas do sequenciamento na plataforma 454 (azul claro) e aquelas preditas pelo Augustus (amarelo), assim como o total de sequências.

Tabela 3. Tabela comparando os dados de produtos gênicos descritos no genoma de *T. rangeli* com os resultados gerados pela etapa de predição gênica e concatenação dos dados no conjunto de proteínas não redundantes (CDS-NR)

	Genoma *T. rangeli*	**CDS – NR**
Conteúdo GC	55 %	55,5 %
Tamanho máximo	13.347	14.949
Tamanho médio	1.377	1.499
Tamanho mínimo	153	153
Bases A	22,12 %	22,00 %
Bases C	25,22 %	25,17 %
Bases G	29,76 %	30,16 %
Bases T	22,90 %	22,41 %
Bases N	< 1 %	0,27 %
Número de Bases	10.274.856	15.749.418
Número de Sequências	7.457	10.506

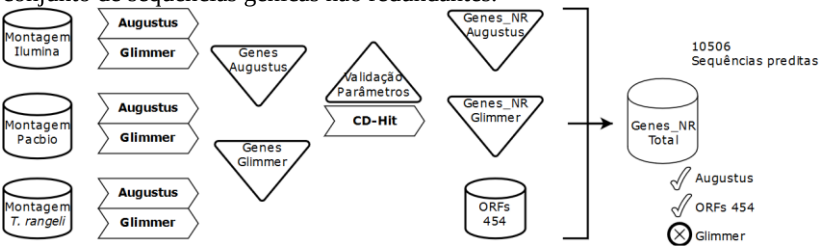
Diante da decisão de excluir os resultados obtidos pelo programa Glimmer, resta a discussão sobre a inclusão de uma ferramenta de predição gênica alternativa para anotação de proteínas hipotéticas. O motivacional principal deste trabalho em utilizar dois preditores gênicos distintos, que utilizam metodologias de predição diferentes, era realizar uma comparação direta entre os seus resultados. Dessa forma, poderíamos decidir qual deles se encaixaria melhor ao nosso conjunto de dado e, posteriormente, confirmar essa decisão quando da utilização deste pipeline para a reanotação de proteínas hipotéticas de outros organismos. A partir da retirada do Glimmer, abre-se espaço para uma nova versão deste pipeline, a qual ainda mantém a ideia de utilizar dois preditores gênicos para realizar esta etapa fundamental no processo de anotação de sequências genômicas.

Alguns possíveis candidatos, conforme recomendado por Yandell e Ence (2012), seriam os preditores gênicos SNAP (KORF, 2004), Geneid (PARRA, 2000) ou mGene (SCHWEIKERT et al., 2009). Os dois primeiros aceitam evidências extrínsecas para construírem seu conjunto

de treino antes da predição, como ESTs e proteínas preditas já descritas. O último utiliza-se de algoritmos de aprendizado de máquinas (*machine learning*), através de máquinas de suporte de vetores de suporte (*Support-Vector Machines* – SVM), para realizar suas predições.

Uma visão esquemática mais específica desta etapa do trabalho pode ser observada na Figura 6, que demonstra todos os arquivos de entrada e o conjunto de dados obtidos no final deste processo, assim como a decisão de manter apenas uma porção dos dados gerados.

Figura 6. Representação gráfica das análises realizadas para a obtenção de um conjunto de sequências gênicas não redundantes.



Tal conjunto foi obtido a partir de sequências preditas pelo preditor Augustus e das sequências oriundas do sequenciamento de *T. rangeli* na plataforma 454, conforme descrito por Stoco e colaboradores (2014).

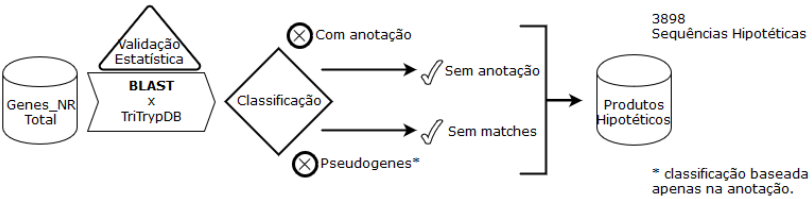
Para as análises subsequentes do pipeline ressaltamos que foram utilizadas as sequências proteicas correspondentes aos CDS-NR, exceto pela etapa de quantificação dos *reads* do transcriptoma nas sequências hipotéticas.

4.2.2 Obtenção do conjunto de proteínas hipotéticas

A partir das sequências proteicas não redundantes obtidas, foi realizada uma etapa de análise de similaridade contra banco de dados de proteínas anotadas com o objetivo de identificar proteínas com funções conhecidas e descritas e, também, aquelas que seriam anotadas como proteínas hipotéticas no genoma de *T. rangeli*. Entretanto, antes que fosse feita a análise com o conjunto de proteínas preditas, se procedeu uma etapa de validação de parâmetros do BLAST+ utilizando dados de sequências já disponíveis no TriTryDB. O objetivo deste teste foi avaliar os parâmetros de identidade e positividade das sequências, para então utilizá-los como valores limítrofes mínimos no conjunto de proteínas CDS-NR.

A Figura 7 demonstra uma visão esquemática de todos os arquivos e programas utilizados, testes realizados e resultados encontrados nesta etapa do trabalho.

Figura 7. Representação esquemática das análises de similaridade realizadas para obtenção das anotações das sequências preditas não redundantes.



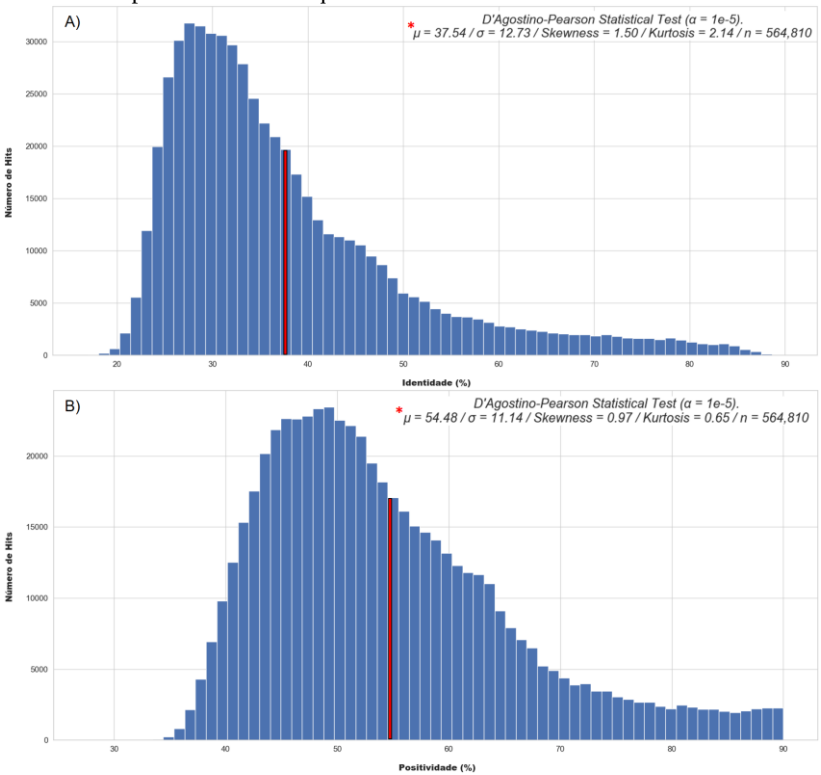
As sequências de anotação hipotéticas e as sequências sem correspondência de similaridade com o banco de dados formaram o conjunto de dados de produtos hipotéticos, com 3.898 sequências no total.

Para tal, foi selecionado um conjunto de sequências dentre as disponíveis no TriTrypDB para simular um conjunto de dados que representasse a diversidade genômica presente dentre os principais grupos de tripanossomatídeos. Seis genomas foram escolhidos (*Leishmania braziliensis* MHOMBR75M2904, *Leishmania major* Friedlin, *Trypanosoma brucei* TREU927, *Trypanosoma congolense* IL3000, *Trypanosoma cruzi* marinkellei B7 e *Trypanosoma cruzi* SylvioX10) e, para garantir que as anotações encontradas seriam apenas de proteínas com função conhecidas, foram removidas deste conjunto de dados as sequências proteicas que continham quaisquer uma das seguintes palavras-chave dentre suas anotações: “fragment”, “hypothetical”, “partial”, “pseudogene” ou “putative”. Em seguida, utilizando os resultados encontrados pelo algoritmo do BLASTp, um script *in-house* buscava em cada um dos resultados encontrados pela análise de similaridade os parâmetros de positividade e identidade que representassem o momento que ocorria uma inflexão de anotação, ou seja, quando a anotação de uma determinada sequência (anotada como *trans-sialidase*, por exemplo) era modificada por uma outra anotação qualquer (*gp63*, por exemplo).

O cálculo estatístico escolhido para analisar os resultados encontrados (n = 564.810) foi o teste de D’Agostino-Pearson, com alfa de 10^{-5}, o qual permite observar valores de *skewness* e *kurtosis* associados à distribuição dos dados. Os parâmetros observados por essa análise de similaridade são apresentados na Figura 8. Foi obtido um valor de identidade de 37,54% e de 54,48% para positividade, o que significa que

é possível transferir a anotação de um resultado de similaridade quando estes valores forem acima deste valor limítrofe.

Figura 8. Figura composta por dois gráficos que demonstram os resultados encontrados pelo teste estatístico de D'Agostino-Pearson (alfa = 10^{-5}) realizados em um conjunto de sequências artificiais para obtenção de parâmetros de identidade e positividade de sequências.



A) Histograma que demonstra a distribuição dos dados para o parâmetro de identidade de sequências, cuja média (μ) do ponto de inflexão é de 37,54. B) Histograma que demonstra a distribuição dos dados para o parâmetro de positividade de sequências, cuja média (μ) do ponto de inflexão é de 54,58. Em ambos os gráficos, a barra vermelha representa a média.

O motivacional desta análise exploratória de dados foi para garantir que todos os resultados encontrados na análise de similaridade fossem confiáveis sendo que, ao mesmo tempo, um ajuste fino dos parâmetros permite que a análise seja o mais fiel o possível ao nosso conjunto de dados.

De posse deste parâmetro, foi realizada a etapa de análise de similaridade com o conjunto de proteínas preditas não redundantes, sendo que os resultados encontrados por essa análise distribuem as 10.506 CDS-NR nas quatro classificações da seguinte maneira: (1) 6.475 sequências "com anotação"; (2) 133 sequências cuja anotação continha a palavra "pseudogene"; (3) 3.740 sequências hipotéticas, cuja anotação continha uma ou mais palavras-chave ("fragment", "hypothetical" ou "partial"); e (4) 158 sequências sem similaridade com o banco de dados utilizado. É importante destacar que a classificação das sequências como pseudogene foi baseada inteiramente na anotação encontrada no TriTrypDB e apesar de existirem descritos na literatura pipelines específicos para a predição e classificação destes pseudogenes (SOLOVYEV et al., 2006; ZHANG et al., 2006), este estudo não objetivou a análise destes dados.

Com base nestes resultados, as sequências classificadas no terceiro e no quarto grupo foram concatenadas em um conjunto de dados chamado de "proteínas hipotéticas", totalizando 3.898 sequências, sendo estas as sequências que foram utilizadas nas demais análises do pipeline. Uma comparação das características gerais destas sequências pode ser visualizada na Tabela 4, onde é possível notar que os atributos como tamanho médio das sequências e conteúdo GC das ORFs preditas do genoma de *T. rangeli*, das CDS-NR e deste conjunto de proteínas hipotéticas, se mantém similares.

Tabela 4. Tabela com os dados dos produtos hipotéticos obtidos das análises de similaridade do pipeline em comparação aos produtos do genoma sequenciado na plataforma 454 e as proteínas preditas não redundantes.

	Genoma *T. rangeli*	**CDS – NR**	**Proteínas Hipotéticas**
Conteúdo GC	55 %	55,5 %	55,8 %
Tamanho máximo	13.347	14.949	14.949
Tamanho médio	1.377	1.499	1.334
Tamanho mínimo	153	153	165
Bases A	22,12 %	22,00 %	21,83 %
Bases C	25,22 %	25,17 %	25,43 %
Bases G	29,76 %	30,16 %	30,34 %
Bases T	22,90 %	22,41 %	22,28 %
Bases N	< 1 %	0,27 %	0,13 %
Número de Bases	10.274.856	15.749.418	5.203.770
Número de Sequências	7.457	10.506	3.898

Considerando que a abordagem utilizada para a anotação dos genes no genoma publicado deste parasito foi baseada em similaridade com bancos de dados (TriTrypDB e SwissProt), método muito similar ao adotado por este trabalho, nota-se que aqui foram encontradas mais CDS do que aquelas descritas por Stoco e colaboradores (2014) na publicação do genoma de *T. rangeli*, assim como uma porcentagem menor de proteínas hipotéticas: 65,6% descritas no genoma do parasito e 37,1% descritas neste trabalho. Um dos principais motivos que podem ser atribuídos para uma diminuição tão significativa no número de proteínas hipotéticas é o fato de que a quantidade de genomas que foram depositados no TriTrypDB aumentou drasticamente desde 2014 até hoje: na versão 6.0 do TriTrypDB, disponibilizada em setembro de 2013, estavam disponíveis genomas de 44 espécies de tripanossomatídeos diferentes, enquanto que na versão atual (43.0) estão disponíveis 89 espécies. A quantidade de sequências depositadas também aumentou, assim como aumentaram os números de sequências melhores anotadas com descrições específicas. Este fato serve para destacar a importância de se utilizar um banco de dados rico e com sequências bem anotadas, que demonstra claramente que através de uma simples análise de similaridade é possível realizar um procedimento de anotação genômica de boa qualidade.

4.2.3 Estudo das anotações e funções dos produtos hipotéticos

A primeira abordagem que foi utilizada para tentar elucidar a possível função das proteínas hipotéticas foi uma abordagem baseada num estudo integrado entre anotações de bancos de dados distintos e possíveis anotações funcionais. A primeira ferramenta de escolha para tal foi o programa InterProScan, que nos permite ter um panorama de anotações de diferentes bancos de dados, assim como classificar as sequências proteicas em superfamílias, mesmo quando não há uma anotação disponível.

Ainda, nas versões mais atuais do InterProScan (versão 5.0+), além das possíveis anotações, o programa nos dá indícios de anotações funcionais daquele determinado produto gênico com base na anotação da Gene Ontology. A anotação funcional do InterProScan (IPR) relaciona as descrições e anotações encontradas pelo programa em diferentes bancos de dados, tais como os domínios proteicos ou as superfamílias proteicas, com as próprias anotações funcionais sugeridas pela Gene Ontology. A Gene Ontology (GENE ONTOLOGY CONSORTIUM, 2015) é uma base de dados de ontologia de anotações, utilizada para facilitar a padronização

das diversas anotações gênicas, no intuito de organizar todo um vocabulário científico e sua multiplicidade de termos e regras dentre os muitos bancos de dados disponíveis. Este vocabulário está organizado em três ontologias principais: (i) componente celular; (ii) processo biológico; e (iii) função molecular, os quais são organizados em uma hierarquia funcional onde os termos podem relacionar-se com termos de diferentes ontologias.

De todas as 3.898 sequências de proteínas hipotéticas analisadas pelo InterProScan, 1.055 encontraram alguma correspondência de descrição com pelo menos um dos bancos de dados ou pelo menos uma anotação funcional, que nos fornece indícios do que poderia se tratar aquela sequência. Na tentativa de avaliar a especificidade dos resultados encontrados pelo InterProScan, buscamos informações sobre os parâmetros estatísticos utilizados na execução do programa em sua documentação (disponível em: https://github.com/ebi-pf-team/interproscan/wiki/FAQ, acessado em 08/11/2018), pois não há como saber se o InterProScan estava sendo executado de maneira muito restritiva ou muito flexível. Como a informação sobre estes parâmetros não foi encontrada, foi decidido que seria realizada uma análise independente que relacionasse dois bancos de dados de interesse, Pfam e CDD, de maneira independente.

Assim, foi realizada a busca por homologias distantes através de modelos matemáticos probabilísticos, como perfis de HMM associados às sequências, através do programa HMMER contra o banco de dados do Pfam encontrou em 478 sequências das 3.898 pelo menos alguma anotação a respeito de seus domínios. O algoritmo do programa HMMER utilizado para fazer a busca por domínios proteicos conservados foi o `hmmscan`, uma vez que estávamos comparando nossas sequências hipotéticas contra um banco de dados de perfis de HMM. Ainda, a utilização de um conjunto de dados curados para produção de perfis de HMM baseados em proteomas de referência acessíveis através do UniProtKB contribui para o elevado grau de confiabilidade dos dados desta análise (FINN et al., 2015). Também é importante destacar que comumente as proteínas recebem seu nome ou são classificadas de acordo com a presença de domínios conhecidos e conservados, muitas vezes podendo ser agrupadas em superfamílias pela presença de certos domínios e alinhamentos de perfis de HMM com outras proteínas (INTERNATIONAL PROTEIN NOMENCLATURE GUIDELINES, 2018).

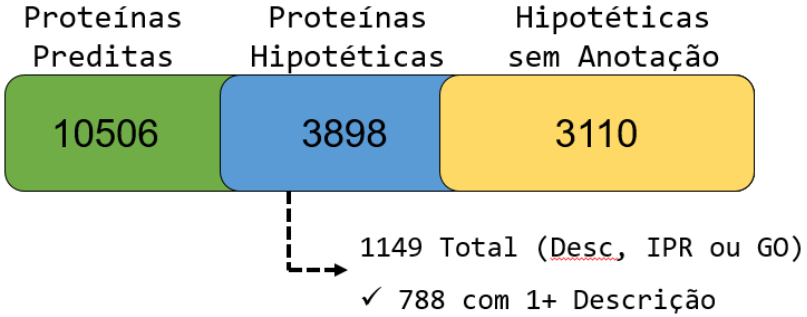
Na última análise realizada nesta etapa foi realizada a avaliação da presença de domínios conservados nas sequências hipotéticas utilizando-

se o programa RPSblast+ com os dados disponíveis no *Conserved Domain Database* – CDD (MARCHLER-BAUER et al., 2012), que é um repositório curado que contém uma coleção de modelos de domínios proteicos conservados ao longo da história evolutiva das superfamílias de proteínas. Os modelos contam com informações tridimensionais e também levam em consideração a variação de resíduos aminoacídicos, tentando encontrar padrões e fazer previsões de como essas mudanças ocorrem dentro de uma mesma superfamília e como elas podem estar relacionadas com as propriedades funcionais das proteínas. Esta análise resultou em 433 das 3.898 sequências de proteínas hipotéticas encontraram alguma correspondência no banco de dados do CDD.

De posse de todas as informações a respeito das descrições e anotações funcionais dessas proteínas, os dados foram organizados por um script *in-house*. Ao final, o número total de sequências que continham ao menos uma anotação ou ao menos uma anotação funcional foi de 1.149 do total de 3.898 sequências hipotéticas. Levando em consideração apenas esse parâmetro como medida para redução do número de proteínas hipotéticas anotadas no genoma de *T. rangeli*, isso significaria uma redução de aproximadamente 29,48%. Porém, como alguns autores sugerem: uma anotação funcional não nos fornece, necessariamente, algum indicativo da função específica daquela proteína, pois muitas vezes os termos ontológicos encontrados dentro de uma hierarquia de alto nível são bastantes generalistas e acabam por abranger uma grande gama de possibilidades de funções (RHEE et al., 2008; THOMAS et al., 2012).

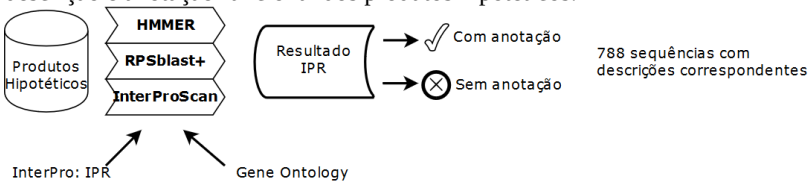
Diante dessa situação, decidimos por considerar apenas os resultados que continham ao menos uma descrição, o que resulta em um número de 788 sequências proteicas com anotação, que poderiam ser anotadas com uma descrição mais específicas do que "proteínas hipotéticas". Esta redução, de um ponto de vista mais conservador, significa uma diminuição de aproximadamente 20,22% na quantidade de proteínas anotadas como hipotéticas no genoma de *T. rangeli* e que seria possível determinar a função de uma proteína hipotética através de estudos de suas descrições e anotação funcional (Figura 9). O número total de proteínas com alguma correspondência de anotação, assim como uma visão esquemática deste resultado estão representados pela Figura 10.

Figura 9. Representação gráfica da quantidade de sequências hipotéticas restantes ao final da análise integrada de descrição e anotação funcional, a partir dos resultados do InterProScan, HMMER e RPSblast+.



Em verde, o número total de produtos gênicos preditos não redundantes. Em azul, o número total de proteínas hipotéticas, oriundas das análises de similaridade. Em amarelo, o número total de proteínas hipotéticas que não encontraram com algum banco de dados para sua descrição ou anotação funcional.

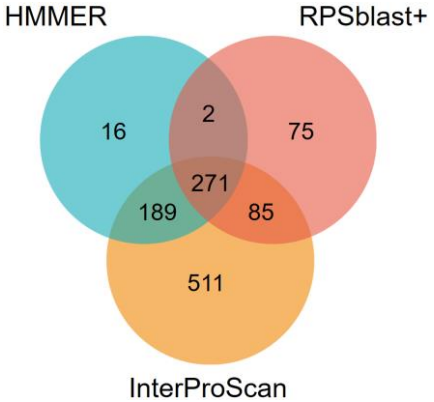
Figura 10. Representação esquemática dos resultados das análises integradas de descrição e anotação funcional dos produtos hipotéticos.



Mostrando o conjunto de dados utilizados, os programas que foram executados e a seleção final das anotações correspondentes. A possibilidade de reanotação de 789 sequências hipotéticas representa uma redução em 20,24% no número de proteínas hipotéticas do genoma de *T. rangeli*.

É importante comentar que o número total de 1.149 sequências descritas nesta etapa do trabalho não é uma somatória direta dos números de sequências encontradas nas análises independentes (1.056 pelo InterProScan, 478 pelo HMMER e 433 pelo RPSblast+), uma vez que tanto os bancos do Pfam como o CDD já estão inclusos nas análises do InterProScan, então há sobreposição de resultados (Figura 11).

Figura 11. Diagrama de Venn que representa a distribuição das 1.149 sequências hipotéticas que obtiveram resultado de correspondência do estudo integrado de anotação e função.



Os subgrupos das sequências estão divididos de acordo com o programa: HMMER (azul claro); RPSblast+ (vermelho claro); e InterProScan (amarelo).

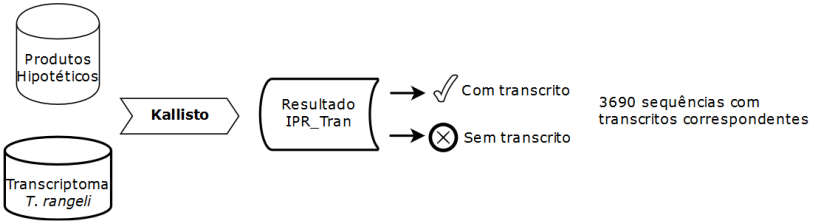
Diante deste panorama, é importante que o usuário deste pipeline pondere se existe algum benefício em realizar as análises de maneira independente, devido ao pouco que é acrescido de informação sobre as suas sequências. A justificativa para a realização aqui, além daquela explanada anteriormente, é justamente explorar a maior gama de possibilidades o possível para nos assegurarmos de que todas as informações disponíveis foram levadas em consideração.

4.2.4 Análises de evidências de expressão

4.2.4.1 A partir de RNAseq – Transcriptoma

A primeira etapa de validação experimental do conjunto de sequências hipotéticas diz respeito a quantificação da abundância de transcritos de *T. rangeli* obtidos através das técnicas de RNAseq (Figura 12). Nesta etapa foram utilizadas sequências nucleotídicas que correspondem aos produtos hipotéticos obtidos neste trabalho. O programa escolhido para realizar esta análise foi o programa Kallisto, que utiliza uma técnica de pseudo-alinhamento para relacionar os *reads* de RNAseq às sequências alvos de maneira eficiente, sem necessidade de alinhamentos individuais de bases, o que reduz significativamente o tempo computacional necessário para realizar análises de quantificação (BRAY et al., 2016).

Figura 12. Representação gráfica dos dados empregados na etapa de avaliação da transcrição dos produtos hipotéticos, o programa utilizado e dos resultados obtidos.



Foram considerados como resultados positivos as 3.690 das 3.898 sequências hipotéticas cuja quantificação mostrou uma quantidade de transcritos por milhão acima de zero.

Como resultados, 3.690 (94,66%) das 3.898 sequências hipotéticas apresentaram resultado positivo para correspondência com transcritos e apenas 208 (5,34%) não apresentaram quaisquer indícios de transcritos correspondentes. Ainda, nossos dados mostram que foram encontrados em média 271 transcritos por milhão (TPM) nas sequências que tiveram seus transcritos quantificados, enquanto que o maior número de transcritos encontrados para uma única sequência hipotética foi de 55.595,6 TPM.

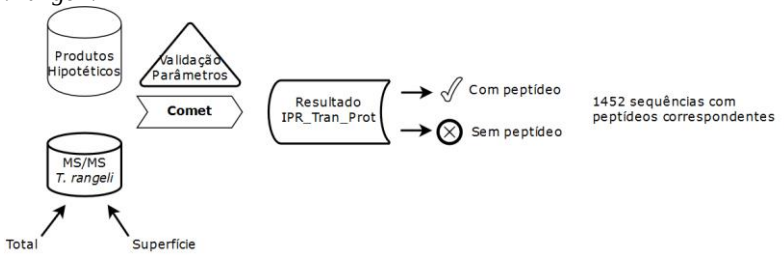
Este elevado grau de resultados positivos para as evidências experimentais de transcritos podem ser atribuídos principalmente a dois fatores: i) um fator biológico associado ao controle da expressão gênica em organismos tripanossomatídeos, que acontece de maneira pós-transcricional através do auxílio de um mini-exon de 39 pares de base (chamado de *spliced leader*), o qual possibilita que estes parasitos estejam transcrevendo seus genes de maneira constitutiva, sem necessariamente traduzi-los (TEIXEIRA, 1998; BEN-DOV; LEVIN; VÁZQUEZ, 2005); e ii) um fator intrínseco à maneira como ocorreu a análise bioinformática onde, através da avaliação da quantificação, foram considerados positivos quaisquer resultados acima de zero, sendo que o menor número de transcritos por milhão para uma determinada sequência foi de 0,000335164 TPM. Desta forma, especificamente para este modelo biológico, a utilização de técnicas para quantificação de produtos de transcrição mostrou-se pouco informativa como forma de validação da existência das proteínas hipotéticas, o que justifica a necessidade de se realizar uma etapa adicional de análises de evidências experimentais com dados proteômicos.

4.2.4.2 A partir de espectrometria de massas – Proteoma

A última análise realizada neste trabalho foi a etapa de investigação dos dados de evidência de expressão de proteínas de *T. rangeli*, cujo papel é importantíssimo tendo em vista o contexto biológico deste organismo. Ainda, segundo a literatura, evidências apontam para uma discrepância entre a quantidade de transcritos produzidos e o número destes que são realmente traduzidos (CASTELLANA et al., 2014), sendo que para organismos modelos não humanos ou camundongos essa discrepância é ainda maior (NESVIZHSKII, 2014). Todos esses fatores corroboram para que a análise de expressão proteica seja uma das etapas mais importantes de validação de proteínas hipotéticas.

O programa utilizado nesta etapa foi o Comet que, basicamente, calcula pontuações (*scores*) entre os peptídeos encontrados através das análises de espectrometria de massa *in tandem* (MS/MS) e as sequências correspondentes no conjunto de dados de sequências proteicas e, então, utiliza a distribuição dessas pontuações para gerar um valor de *e* (ENG; JAHAN; HOOPMANN, 2012). É importante destacar que os parâmetros utilizados na execução do Comet foram bastante restritivos, justamente para garantir um alto grau de confiabilidade dos resultados encontrados. Da mesma forma, o tratamento de bioinformática que avaliou os resultados encontrados também foi bastante restritivo, como uma maneira de assegurar a qualidade da verificação experimental. Na Figura 13 é possível observar uma representação de todos os dados utilizados e dos resultados obtidos nesta etapa.

Figura 13. Representação esquemática das análises de correspondência entre os produtos hipotéticos e proteínas do parasito. Foram utilizadas as sequências de proteínas hipotéticas e dados de espectrometria de massas total e de superfície de *T. rangeli*.

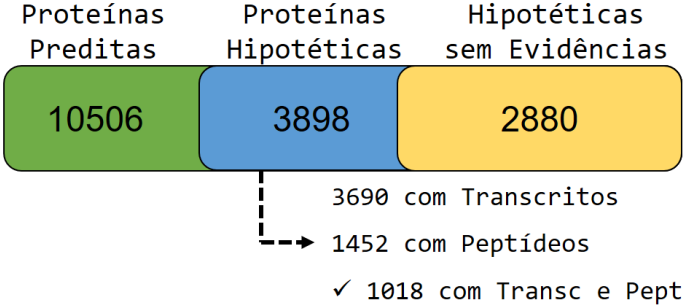


No total, 1.452 das 3.898 sequências hipotéticas encontraram correspondência com ao menos 2 peptídeos diferentes do parasito, o que sugere a possibilidade de que essas proteínas hipotéticas estejam sendo expressas ao longo do ciclo de vida deste parasito.

Para esta etapa foram utilizadas 91 análises de espectrometria de massas, oriundas de amostras biológicas distintas (proteínas solúveis totais e de superfície) e de diferentes fases do ciclo de vida do parasito. Nossos resultados indicam que 1.452 (37,25%) das 3.898 sequências hipotéticas apresentaram pelo menos dois peptídeos nas análises de espectrometria de massas (FDR – *False Discovery Rate* – de 0,24 ou 24%). As outras sequências restantes não passaram nos filtros de avaliação ou não tiveram correspondência no banco de dados de peptídeos e não foram consideradas.

Levando em consideração uma perspectiva mais completa das evidências experimentais das sequências hipotéticas, tomou-se a decisão de agrupar os dois conjuntos de dados. Dessa forma, foram consideradas apenas aquelas sequências que apresentavam ambas evidências de expressão, totalizando 1.018 sequências de proteínas hipotéticas potencialmente expressas (Figura 14).

Figura 14. Representação gráfica da quantidade de sequências hipotéticas restantes ao final da análise de evidência de expressão, considerando os resultados de correspondência obtidas através dos programas Kallisto e Comet.



Em verde, o número total de produtos gênicos preditos não redundantes. Em azul, o número total de proteínas hipotéticas, oriundas das análises de similaridade. Em amarelo, o número total de proteínas hipotéticas que não apresentam evidência de transcrição ou expressão.

É importante comentar que essa avaliação não leva em consideração a presença de dados de descrição ou anotação dessas sequências, uma vez que estamos avaliando apenas a possibilidade de expressão das mesmas. Este resultado representaria uma redução de aproximadamente 26,16% na quantidade de proteínas anotadas como hipotéticas no genoma de *T. rangeli*, além de levantar uma discussão interessante a respeito da classificação de proteínas sem anotações específicas, mas com significativos dados experimentais de expressão.

Wagner (2012) já havia sugerido a utilização de uma nova nomenclatura para a classificação dessas proteínas, que seriam chamadas de “proteínas de função desconhecida” ao invés de “proteínas hipotéticas”, justamente por conta do respaldo experimental que apoia a existência desses produtos gênicos. Um dos principais motivos que justificaria a utilização desta nomenclatura é o fato de que proteínas que contém as descrições “hypothetical”, “probable”, “predicted” ou até “putative” são ignoradas das análises experimentais na grande maioria dos estudos (SCHNOES et al., 2009). Por outro lado, de acordo com Klimke et al. (2011) e a *International Protein Nomenclature Guidelines* (2018), a utilização dessas nomenclaturas alternativas apenas aumentaria ainda mais a heterogeneidade das anotações proteicas e, por causa disso, devem ser evitadas para este tipo de anotação funcional de genes.

4.2.5 Classificação dos produtos hipotéticos no pipeline com base no modelo de *Trypanosoma rangeli*

A partir da interpretação dos resultados encontrados neste trabalho, decidimos criar oito categorias para classificarmos as sequências proteicas hipotéticas preditas com base em três critérios de avaliação: (1) possíveis descrições de anotação nos bancos de dados utilizados; (2) correspondência com evidências de transcrição; e (3) presença de peptídeos detectados nos bancos de dados de espectrometria de massas. O número total de cada classe e o quanto essa classe representa considerando o total de 3.898 sequências proteicas hipotéticas preditas, estão demonstrados na Tabela 5.

Tabela 5. Tabela de classificação das 3.898 sequências hipotéticas descritas neste trabalho.

	Possível Descrição	Evidência de Transcrição	Evidência de Expressão	Número de Sequências	Porcentagem (%)
Classe 1	-	-	-	121	3,18
Classe 2	✓	-	-	22	0,56
Classe 3	-	✓	-	1.912	49,05
Classe 4	-	-	✓	56	1,44
Classe 5	✓	✓	-	388	9,95
Classe 6	✓	-	✓	6	0,15
Classe 7	-	✓	✓	1.018	26,12
Classe 8	✓	✓	✓	372	9,54

Legenda: - não atendeu ao critério de classificação. ✓ atendeu ao critério de classificação.

As sequências que integram a classe 1 poderiam ser consideradas como proteínas realmente hipotéticas, pois não há nenhum indício de sua existência a não ser a predição *in silico*. Contudo, a classe 1 e a classe 2 também poderiam ser consideradas como artefatos computacionais dos preditores gênicos, pois não apresentaram evidências de expressão. Cabe ressaltar que até os melhores preditores gênicos raramente conseguem ultrapassar uma acurácia de 80% em suas predições (YANDELL; ENCE, 2012). Sendo assim, a predição destes genes poderia ser revisada para avaliar se estes realmente seriam artefatos de predição, dessa forma evitando que estes genes fossem anotados como "hipotéticos". A anotação incorreta de um gene pode possivelmente refletir em uma transferência de anotação errônea para um outro genoma e, segundo Roberts (2004), estima-se que essas anotações equivocadas podem representar até mais da metade das potenciais regiões codificantes de um novo genoma.

Nota-se a grande quantidade de sequências que pertencem à classe 3, que representa 49,05% de todas as sequências hipotéticas, as quais apresentam apenas evidência de transcrição. Mais uma vez, torna-se

irrefutável o aspecto biológico da expressão constitutiva das unidades funcionais do genoma de *T. rangeli*. Curiosamente, 26,12% das sequências pertencem à classe 7, a qual apresenta sequências que não têm correlação com os bancos de dados utilizados, mas sua expressão está sustentada por evidências experimentais, tanto de transcrição como de expressão proteica. Isto pode indicar que estas sequências gênicas são exclusivas de *T. rangeli*, também chamados de genes órfãos. Genes órfãos são classificados como genes que não apresentam similaridade com nenhum genoma disponível nas bases de dados existentes (TAUTZ; DOMAZET-LOŁO, 2011). Estima-se que em torno de 10 a 20% dos genes preditos em um genoma são genes órfãos (FUKUCHI; NISHIKAWA, 2004), além disto, estudos demonstram que a maioria dos genes órfãos apresentam evidência de expressão, tanto no âmbito transcricional quanto proteico (PRABH; RÖDELSPERGER, 2016). Considerando este critério de definição de genes órfãos e o total de sequências gênicas preditas não redundantes (10.506), podemos estimar que o genoma de *T. rangeli* apresenta em torno de 9,68% possíveis genes órfãos, um valor similar ao observado em outros genomas de tripanossomatídeos (CALLEJAS-HERNÁNDEZ; GIRONÈS; FRESNO, 2018; CALLEJAS-HERNÁNDEZ et al., 2018).

Finalmente, as proteínas que fazem parte da classe 8 poderiam ser consideradas para a reanotação de proteínas hipotéticas de *T. rangeli*, uma vez que ela é composta por sequências que apresentam descrições disponíveis nos bancos de dados utilizados, evidência de transcrição e evidência de expressão proteica. Essas 372 sequências representam 9,54% de todas as proteínas hipotéticas descritas neste trabalho, sendo estas as que apresentam maior respaldo para uma reanotação confiável. Caberia agora realizar um processo minucioso de averiguação destes dados e submissão de uma nova descrição para estes produtos gênicos.

4.2.6 Considerações

Os programas utilizados para comparação de sequências neste pipeline trabalham utilizando técnicas de alinhamento de sequências, que por natureza apresenta algumas limitações, principalmente quando se comparam sequências que têm baixa porcentagem de similaridade e entram na "zona de penumbra" (*twilight zone*), que representa a faixa entre 20 e 35% de similaridade entre as sequências (BLAKE; COHEN, 2001; PEARSON, 2015). Isso influencia diretamente na confiabilidade dos dados de anotação e descrição das sequências, principalmente quando

o alinhamento acontece entre duas sequências distantes filogeneticamente, como no caso das homologias distantes.

Existem na literatura relatos de programas que utilizam uma abordagem do tipo sem-alinhamento (*alignment-free*) (WOOD; SALZBERG, 2014; OUNIT; LONARDI, 2016), os quais podem ser ferramentas bastantes úteis durante os processos de análises de similaridade de sequências por dois motivos principais: (i) eles permitem realizar análises de similaridade entre sequências sem a necessidade de alinhamentos individuais entre os pares de bases, o que resulta em uma análise de similaridade que inclui tanto a possibilidade de se trabalhar com sequências de organismos filogeneticamente aparentados quanto a possibilidade de se trabalhar com sequências que apresentem características de homologia distantes; e (ii) é um uma metodologia adicional para se comparar os resultados obtidos de uma análise de similaridade típica utilizando um algoritmo conhecido, como aqui no caso o algoritmo do BLAST+. Os programas que realizam essa abordagem sem-alinhamento já têm sua eficiência comprovada, sendo utilizados principalmente em pipelines de anotação de dados de metagenomas, que aplicam este método computacional por conta de sua rapidez (ZIELEZINSKI et al., 2017).

De maneira geral, a utilização de um pipeline que realiza diferentes abordagens computacionais tipicamente proteogenômicas para encontrar possíveis anotações de proteínas hipotéticas se mostrou uma boa ferramenta, possivelmente também para revisar as anotações de proteínas com funções já bem descritas. O único fator que impediu um processo de anotação completamente automático foi justamente o fato da nossa ferramenta implementar informações de diferentes fontes, que muitas vezes apresentam resultados heterogêneos. Dessa forma, para completar o processo de anotação, seria necessária uma etapa manual de avaliação de cada uma das possíveis descrições e anotações funcionais de cada um dos produtos gênicos em questão, levando em consideração todo o aspecto biológico e *know-how* sobre este organismo, o qual caracteriza um processo de curagem manual dos dados.

Apesar de não ser possível tomar essa decisão de maneira automática, nossos resultados incluem todas as possíveis descrições e anotações funcionais encontradas de maneira organizada para tornar este processo manual mais rápido e menos custoso em questão de tempo (Figura 15). Por fim, estão apresentadas as possíveis superfamílias proteicas das quais as proteínas hipotéticas pertencem, para melhor guiar eventuais estudos que busquem caracterizar experimentalmente essas proteínas.

Figura 15. Exemplo da tabela de informações e classificação final das proteínas hipotéticas analisadas por este pipeline.

Query	Protein Superfamily	Descrição	IPR	GO	TPM	Prot
AUG_llu_g4346.t1_AUG_Pac	TerB-like (SSF158682)	cd07311, terB_like_1, tellurium resistance te	InterPro:IPR029024	None	161.164	No peptide evidence
AUG_llu_g4347.t1_AUG_Pac	None	None	None	None	146.679	No peptide evidence
AUG_llu_g4348.t1_AUG_Pac	AH/BAR domain superfamily (SSF1(	None	None	None	1429.41	No peptide evidence
AUG_llu_g4349.t1_AUG_Pac	None	None	None	None	438.908	Número total de Peptideos: 4 Número de peptideos não redundantes: 3 EQYTQSGRNK, QESSFMGFLR, RMGVQPTP IONS_MATCHED: 28
AUG_llu_g4353.t1_AUG_Pac	None	None	None	None	334.881	Número total de Peptideos: 11 Número de peptideos não redundantes: 3 EGDARCGPALR, HVYYYEDRAEQGHK, IEGL IONS_MATCHED: 105
AUG_llu_g4358.t1_AUG_Pac	None	None	None	None	106.343	No peptide evidence
AUG_llu_g4359.t1	Alkaline-phosphatase-like, core dom	cd16021, ALP_like, uncharacterized Alkaline Protein of unknown function (DUF229)	InterPro:IPR004245	None	158.94	Número total de Peptideos: 3 Número de peptideos não redundantes: 3 IERCSDLGALNSK, VDFCGLEGLK, YGMNET IONS_MATCHED: 23
AUG_llu_g4361.t1_AUG_Pac_	1 (SSF69635)	PDZ_serine_protease; cd00987, PDZ_serine_protease, PDZ doma Tir chaperone protein (CesT) family, PDZ domain	InterPro:IPR010261 InterPro:IPR001478	GO:0005737 GO:0050708 GO:0005515	133.363	Número total de Peptideos: 4 Número de peptideos não redundantes: 3 LQDDVEAAER, GTQEQSEIR, RAADLATR IONS_MATCHED: 28
AUG_llu_g4360.t1_AUG_Pac	CH domain superfamily (SSF47576)	None	InterPro:IPR036872 InterPro:IPR027328	GO:0008017	193.906	No peptide evidence

O arquivo completo está disponível para download em: https://drive.google.com/file/d/1nwohsgRysHqrtzeZMAwF0j66kolp2Nj1/view?usp=sharing

5 CONCLUSÕES

- A escolha de uma ferramenta robusta para predizer as unidades funcionais do genoma de um organismo foi uma etapa crucial durante o desenvolvimento deste trabalho. Neste caso, o programa Augustus provou-se um preditor gênico de excelência em função da baixa quantidade de artefatos gerados.
- A etapa de anotação funcional através do programa InterProScan foi essencial aos resultados obtidos por este trabalho, tanto pela possibilidade de se avaliar uma mesma sequência sob diferentes pontos de vistas (bancos de dados), quanto pela capacidade de se realizar uma anotação funcional que integrasse os termos da ontologia da Gene Ontology de maneira rápida e eficiente.
- Realizar de maneira independente as análises de homologia distante e busca por domínios proteicos conservados adicionou pouco valor às informações que já haviam sido relacionadas através do estudo integrado na plataforma InterProScan.
- Estudar a correspondência de evidências de transcrição nos produtos hipotéticos de *T. rangeli* através de análises quantitativas mostrou-se uma abordagem pouco informativa, pois a grande quantidade de resultados positivos não nos possibilitou inferir a existência das sequências sem realizar uma outra análise confirmatória.
- As análises de evidência de expressão proteica das sequências hipotéticas forneceram um respaldo experimental muito necessário para justificar os resultados encontrados das análises de transcrição, servindo como peça chave para a possibilidade de reanotação dessas sequências.
- Considerando dados genômicos, transcriptômicos e proteômicos nosso pipeline demonstrou que uma série de análises *in silico* possibilita uma redução de 9,54% no número de proteínas hipotéticas de *T. rangeli*.

6 PERSPECTIVAS

Com esta primeira versão do pipeline ajustado para uma espécie de tripanossomatídeo, é do nosso interesse testá-lo com outras espécies filogeneticamente aparentadas destes parasitos (*Trypanosona cruzi*, *Trypanosoma brucei*, etc) para verificar e validar a sua funcionalidade. Bem como, utilizar outro organismo modelo cujo genoma já está melhor anotado para fazermos um comparativo das análises.

Além disto, implementaremos uma nova ferramenta de predição gênica no pipeline, para possibilitar a comparação entre os genes preditos nos dois programas, uma vez que o Augustus já é um preditor que utiliza dados de expressão e validação de proteínas, daríamos preferência para um programa que utilize métodos de aprendizado de máquina. Neste cenário, o programa mGene seria um bom candidato, pela utilização de técnicas de SVM que realizariam a predição gênica.

Também pretendemos desenvolver um algoritmo de aprendizado de máquinas, que utilizaria técnicas de mineração de texto para avaliar a heterogeneidade das possíveis descrições e anotações funcionais disponíveis em bancos de dados, com o objetivo de realizar automaticamente o processo final de curagem para a anotação das sequências de proteínas hipotéticas.

7 REFEREÊNCIAS

AFCHAIN, D. et al. Antigenic Make-Up of *Trypanosoma cruzi* Culture Forms: Identification of a Specific Component. **The Journal Of Parasitology**, v. 65, n. 4, p.507-515, ago. 1979. JSTOR.

ASLETT, M. et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. **Nucleic Acids Research**, v. 38, n. 1, p.457-462, 20 out. 2009. Oxford University Press (OUP).

BEN-DOV, C. P.; LEVIN, M. J.; VÁZQUEZ, M. P. Analysis of the highly efficient pre-mRNA processing region HX1 of *Trypanosoma cruzi*. **Molecular And Biochemical Parasitology**, v. 140, n. 1, p.97-105, mar. 2005. Elsevier BV.

BLAKE, J. D.; COHEN, F. E. Pairwise sequence alignment below the twilight zone. **Journal Of Molecular Biology**, v. 307, n. 2, p.721-735, mar. 2001. Elsevier BV.

BRAY, N. L et al. Near-optimal probabilistic RNA-seq quantification. **Nature Biotechnology**, v. 34, n. 5, p.525-527, 4 abr. 2016. Springer Science and Business Media LLC.

BRUCE, D. Classification of the African Trypanosomes pathogenic to man and domestic animals. **Transactions Of The Royal Society Of Tropical Medicine And Hygiene**, v. 8, n. 1, p.1-22, nov. 1914.

CALLEJAS-HERNÁNDEZ, F.; GIRONÈS, N.; FRESNO, M. Genome Sequence of *Trypanosoma cruzi* Strain Bug2148. **Genome Announcements**, v. 6, n. 3, p.1-2, 18 jan. 2018. American Society for Microbiology.

CALLEJAS-HERNÁNDEZ, F. et al. Genomic assemblies of newly sequenced *Trypanosoma cruzi* strains reveal new genomic expansion and greater complexity. **Scientific Reports**, v. 8, n. 1, p.1-13, 2 out. 2018. Springer Nature.

CAMACHO, C. et al. BLAST+: architecture and applications. **Bmc Bioinformatics**, v. 10, n. 1, p.421-429, 2009. Springer Nature.

CASTELLANA, N. E. et al. An Automated Proteogenomic Method Uses Mass Spectrometry to Reveal Novel Genes in *Zea mays*. **Molecular & Cellular Proteomics**, v. 13, n. 1, p.157-167, 18 out. 2013. American Society for Biochemistry & Molecular Biology (ASBMB).

CHAGAS, C. Neue Trypanosomen: Vorläufige mitteilung. **Archiv für Schiffs- und Tropen-Hygiene**, Leipzig, n.13, p.120-122. 1909a.

CHEN, J. et al. A comprehensive review and comparison of different computational methods for protein remote homology detection. **Briefings In Bioinformatics**, v. 19, n. 2, p.231-244, 13 nov. 2016. Oxford University Press (OUP).

D'ALESSANDRO, A. Biology of *Trypanosoma* (Herpetosoma) *rangeli* Tejera, 1920. In: LUMSDEN, W. H. R.; EVANS, D. A. (Ed.). **Biology of the kinetoplastida**. London: London Academic, v. 3, p.327-403. 1976.

D'ALESSANDRO, A.; SARAVIA, N. G. *Trypanosoma rangeli*. In: Gilles, H. M. **Protozoal Diseases**, Oxford University Press, Oxford. p.398-412. 1999.

D'ARGENIO, V.; SALVATORE, F. The role of the gut microbiome in the healthy adult status. **Clinica Chimica Acta**, v. 451, p.97-102, dez. 2015. Elsevier BV.

DAVILA, A. M. R. et al. GARSA: genomic analysis resources for sequence annotation. **Bioinformatics**, v. 21, n. 23, p.4302-4303, 6 out. 2005. Oxford University Press (OUP).

DELCHER, A. L. et al. Identifying bacterial genes and endosymbiont DNA with Glimmer. **Bioinformatics**, v. 23, n. 6, p.673-679, 19 jan. 2007. Oxford University Press (OUP).

DOLLET, M. Plant Diseases Caused by Flagellate Protozoa (*Phytomonas*). **Annual Review Of Phytopathology**, v. 22, n. 1, p.115-132, set. 1984. Annual Reviews.

EDDY, S. HMMER – Biological sequence analysis using profile hidden Markov models Version 3.1b2. http://hmmer.org/. **Howard Hughes Medical Institute and Department of Genetics Washing University Scholl of Medicine**, St. Louis, USA. 2015.

ENG, J. K.; JAHAN, T. A.; HOOPMANN, M. R. Comet: An open-source MS/MS sequence database search tool. **Proteomics**, v. 13, n. 1, p.22-24, 4 dez. 2012. Wiley.

EL-SAYED, N. M. et al. Comparative Genomics of Trypanosomatid Parasitic Protozoa. **Science**, v. 309, n. 5733, p.404-409, 15 jul. 2005. American Association for the Advancement of Science (AAAS).

FERREIRA, K. A. M. et al. Genome Survey Sequence Analysis and Identification of Homologs of Major Surface Protease (gp63) Genes in *Trypanosoma rangeli*. **Vector-borne And Zoonotic Diseases**, v. 10, n. 9, p.847-853, nov. 2010. Mary Ann Liebert Inc.

FICKETT, J. W. Finding genes by computer: the state of the art. **Trends In Genetics**. v. 12, n. 8, p.316-320, ago. 1996. Elsevier BV.

FINN, R. D. et al. The Pfam protein families database: towards a more sustainable future. **Nucleic Acids Research**, v. 44, n. 1, p.279-285, 15 dez. 2015. Oxford University Press (OUP).

FUKUCHI, S. NISHIKAWA, K. Estimation of the Number of Authentic Orphan Genes in Bacterial Genomes. **Dna Research**, v. 11, n. 4, p.219-231, 1 jan. 2004. Oxford University Press (OUP).

GALLONE, B. et al. Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts. **Cell**, v. 166, n. 6, p.1397-1410, set. 2016. Elsevier BV.

GALPERIN, M. Y. Conserved 'Hypothetical' Proteins: New Hints and New Puzzles. **Comparative And Functional Genomics**, v. 2, n. 1, p.14-18, 2001. Hindawi Limited.

GAO, L. et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. **Nature Genetics**, 13 mai. 2019. Springer Science and Business Media LLC.

GAUTHIER, J. et al. A brief history of bioinformatics. **Briefings In Bioinformatics**, p.1-16, 3 ago. 2018. Oxford University Press (OUP).

GENE ONTOLOGY CONSORTIUM. The Gene Ontology: going forward. **Nucleic Acids Research**, v. 43, p.1049-1056. 2015.

GOVERNMENT OF NEW SOUTHERN WALES. **Genomic pipelines**. 2018. Disponível em: <https://www.genetics.edu.au/genomic/analysis/Genomic-pipelines>. Acesso em: 04 de maio de 2019.

GUHL, F.; VALLEJO, G. A. *Trypanosoma* (*Herpetosoma*) *rangeli* Tejera, 1920: an updated review. **Memórias do Instituto Oswaldo Cruz**, v. 98, n. 4, p.435-442, jun. 2003. FapUNIFESP (SciELO).

GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nature Biotechnology**, v. 29, n. 7, p.644-652, 15 mai. 2011. Springer Nature.

GRISARD, E. C.; STEINDEL, M. *Trypanosoma* (*Herpetosoma*) *rangeli*. In: NEVES, D. P. (ed.) **Parasitologia Humana**, 2004.

GRISARD, E. C. et al. Transcriptomic analyses of the avirulent protozoan parasite *Trypanosoma rangeli*. **Molecular And Biochemical Parasitology**, v. 174, n. 1, p.18-25, nov. 2010. Elsevier BV.

HAGEN, J. B. The origins of bioinformatics. **Nature Reviews Genetics**, v. 1, n. 3, p.231-236, dez. 2000. Springer Nature.

HARROW, J. et al. GENCODE: The reference human genome annotation for The ENCODE Project. **Genome Research**, v. 22, n. 9, p.1760-1774, 1 set. 2012. Cold Spring Harbor Laboratory.

HOLT, C.; YANDELL, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. **Bmc Bioinformatics**, v. 12, n. 1, dez. 2011. Springer Nature.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p.860-921, fev. 2001. Springer Nature.

INTERNATIONAL PROTEIN NOMENCLATURE GUIDELINES. Disponível em <https://www.ncbi.nlm.nih.gov/genome/doc/internatprot_nomenguide/> Acessado em: 09 abr. 2019.

JONES, P. et al. InterProScan 5: genome-scale protein function classification. **Bioinformatics**, v. 30, n. 9, p.1236-1240, 21 jan. 2014. Oxford University Press (OUP).

KANEHISA, M.; SATO, Y.; MORISHIMA, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. **Journal Of Molecular Biology**, v. 428, n. 4, p.726-731, fev. 2016. Elsevier BV.

KLIMKE, W. et al. Solving the Problem: Genome Annotation Standards before the Data Deluge. **Standards In Genomic Sciences**, v. 5, n. 1, p.168-193, 1 out. 2011. Springer Nature.

KOONIN, E. V. Orthologs, Paralogs, and Evolutionary Genomics. **Annual Review Of Genetics**, v. 39, n. 1, p.309-338, dez. 2005. Annual Reviews.

KORF, I. Gene finding in novel genomes. **Bmc Bioinformatics**, v. 5, n. 1, p.59-67, 2004. Springer Nature.

KROGH, A. et al. Hidden Markov Models in Computational Biology. **Journal Of Molecular Biology**, v. 235, n. 5, p.1501-1531, fev. 1994. Elsevier BV.

KUMAR, D. et al. Integrated Transcriptomic-Proteomic Analysis Using a Proteogenomic Workflow Refines Rat Genome Annotation. **Molecular & Cellular Proteomics**, v. 15, n. 1, p.329-339, 11 nov. 2015. American Society for Biochemistry & Molecular Biology (ASBMB).

KUMAR, D. et al. Integrating transcriptome and proteome profiling: Strategies and applications. **Proteomics**, v. 16, n. 19, p.2533-2544, 25 ago. 2016. Wiley.

LI, W.; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, v. 22, n. 13, p.1658-1659, 26 mai. 2006. Oxford University Press (OUP).

LEDERGERBER, C.; DESSIMOZ, C. Base-calling for next-generation sequencing platforms. **Briefings In Bioinformatics**, v. 12, n. 5, p.489-497, 18 jan. 2011. Oxford University Press (OUP).

LEISHMAN, W. B. On the possibility of the occurance of trypanosomiasis in India. **British Medical Journal**. v. 2213, n. 1, p. 1252-1254. 1903.

LUBEC, G. et al. Searching for hypothetical proteins: Theory and practice based upon original data and literature. **Progress In Neurobiology**, v. 77, n. 1-2, p.90-127, set. 2005. Elsevier BV.

LÜCKEMEYER, D. D. Avaliação do perfil proteico de *Trypanosoma rangeli* durante o processo de diferenciação celular in vitro. 2014. Tese de Doutorado (Programa de Pós-Graduação em Biotecnologia e Biociências). Universidade Federal de Santa Catarina, Florianópolis.

LUKEŁ, J. et al. Evolution of parasitism in kinetoplastid flagellates. **Molecular And Biochemical Parasitology**, v. 195, n. 2, p.115-122, jul. 2014. Elsevier BV.

LUSCOMBE, N. M.; GREENBAUM, D.; GERSTEIN, M. What is bioinformatics? An introduction and overview. **Yearbook Of Medical Informatics**, v. 10, n. 01, p.83-100, ago. 2001. Georg Thieme Verlag KG.

MARCHLER-BAUER, A. et al. CDD: conserved domains and protein three-dimensional structure. **Nucleic Acids Research**, v. 41, n. 1, p.348-352, 28 nov. 2012. Oxford University Press (OUP).

MAXAM, A. M.; GILBERT, W. A new method for sequencing DNA. **Proceedings Of The National Academy Of Sciences**, v. 74, n. 2, p.560-564, 1 fev. 1977. Proceedings of the National Academy of Sciences.

MORIYA, Y. et al. KAAS: an automatic genome annotation and pathway reconstruction server. **Nucleic Acids Research**, v. 35, p.182-185, 8 mai. 2007. Oxford University Press (OUP).

MORAES, M. H de et al. Different serological cross-reactivity of *Trypanosoma rangeli* forms in *Trypanosoma cruzi*-infected patients sera. **Parasites & Vectors**, v. 1, n. 1, p.20-29, 2008. Springer Nature.

NESVIZHSKII, A. I. Proteogenomics: concepts, applications and computational strategies. **Nature Methods**, v. 11, n. 11, p.1114-1125, 30 out. 2014. Springer Nature.

NEVES, D. P. et al. **Parasitologia Humana**. 12ª Edição. ed. São Paulo: Atheneu, 2011.

PAEZ-ESPINO, D. et al. Uncovering Earth's virome. **Nature**, v. 536, n. 7617, p.425-430, ago. 2016. Springer Nature.

PARRA, G. GeneID in *Drosophila*. **Genome Research**, v. 10, n. 4, p.511-515, 1 abr. 2000. Cold Spring Harbor Laboratory.

PARRA, G.; BRADNAM, K.; KORF, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. **Bioinformatics**, v. 23, n. 9, p.1061-1067, 1 mar. 2007. Oxford University Press (OUP).

PEARSON, W. R. Protein Function Prediction: Problems and Pitfalls. **Current Protocols In Bioinformatics**, p.4121-4128, 3 set. 2015. John Wiley & Sons, Inc.

PEDERSEN, H. K. et al. Human gut microbes impact host serum metabolome and insulin sensitivity. **Nature**, v. 535, n. 7612, p.376-381, jul. 2016. Springer Nature.

PRABH, N.; RÖDELSPERGER, C. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? **Bmc Bioinformatics**, v. 17, n. 1, p.1-13, 31 mai. 2016. Springer Nature.

OUNIT, R.; LONARDI, S. Higher classification sensitivity of short metagenomic reads with CLARK- S. **Bioinformatics**, v. 32, n. 24, p.3823-3825, 18 ago. 2016. Oxford University Press (OUP).

RHEE, S. Y. et al. Use and misuse of the gene ontology annotations. **Nature Reviews Genetics**, v. 9, n. 7, p.509-515, 13 mai. 2008. Springer Science and Business Media LLC.

RICE, P.; LONGDEN, I.; BLEASBY, A. EMBOSS: The European Molecular Biology Open Software Suite. **Trends In Genetics**, v. 16, n. 6, p.276-277, jun. 2000. Elsevier BV.

ROBERTS, R. J. Identifying Protein Function—A Call for Community Action. **Plos Biology**, v. 2, n. 3, p.42-43, 16 mar. 2004. Public Library of Science (PLoS).

ROMANO, P.; MARRA, D.; MILANESI, L. Web services and workflow management for biological resources. **Bmc Bioinformatics**, v. 6, n. 4, p.24-32, 2005. Springer Nature.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings Of The National Academy Of Sciences**, v. 74, n. 12, p.5463-5467, 1 dez. 1977. Proceedings of the National Academy of Sciences.

SCHNOES, A. M. et al. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. **Plos Computational Biology**, v. 5, n. 12, p.1000605-1000617, 11 dez. 2009. Public Library of Science (PLoS).

SCHWARZE, K. et al. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. **Genetics In Medicine**, v. 20, n. 10, p.1122-1130, 15 fev. 2018. Springer Science and Business Media LLC.

SCHWEIKERT, G. et al. MGene: Accurate SVM-based gene finding with an application to nematode genomes. **Genome Research**, v. 19, n. 11, p.2133-2143, 29 jun. 2009. Cold Spring Harbor Laboratory.

SEEMANN, T. Prokka: rapid prokaryotic genome annotation. **Bioinformatics**, v. 30, n. 14, p.2068-2069, 18 mar. 2014. Oxford University Press (OUP).

SHEYNKMAN, G. M. et al. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. **Bmc Genomics**, v. 15, n. 1, p.703-711, 2014. Springer Nature.

SNOEIJER, C. et al. *Trypanosoma rangeli* Transcriptome Project: Generation and analysis of expressed sequence tags. **Kinetoplastid Biology And Disease**, v. 3, n. 1, p.1-4, 2004. Springer Nature.

SOLOVYEV, V. et al. Automatic annotation of eukaryotic genes, pseudogenes and promoters. **Genome Biology**, v. 7, n. 1, p.10-21, 2006. Springer Nature.

SOUALMIA, L. F.; LECROQ, T. From Genome Sequencing to Bedside. **Yearbook Of Medical Informatics**, v. 22, n. 01, p.175-177, ago. 2013. Georg Thieme Verlag KG.

STANKE, M.; MORGENSTERN, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. **Nucleic Acids Research**, v. 33, p.465-467, 1 jul. 2005. Oxford University Press (OUP).

STEIN, L. Genome Annotation: from sequence to biology. **Nature Reviews Genetics**, v. 2, n. 7, p.493-503, jul. 2001. Springer Nature.

STEVERDING, D. The history of leishmaniasis. **Parasites & Vectors**, v. 10, n. 1, 15 fev. 2017. Springer Nature.

STOCO, P. H. et al. Genome of the Avirulent Human-Infective Trypanosome—*Trypanosoma rangeli*. **Plos Neglected Tropical Diseases**, v. 8, n. 9, p.3176-3192, 18 set. 2014. Public Library of Science (PLoS).

SUN, Z. et al. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. **Nature Communications**, v. 6, n. 1, 29 set. 2015. Springer Science and Business Media LLC.

TATUSOVA, T. et al. NCBI prokaryotic genome annotation pipeline. **Nucleic Acids Research**, v. 44, n. 14, p.6614-6624, 24 jun. 2016. Oxford University Press (OUP).

TAUTZ, D.; DOMAZET-LOŁO, T. The evolutionary origin of orphan genes. **Nature Reviews Genetics**, v. 12, n. 10, p.692-702, 31 ago. 2011. Springer Science and Business Media LLC.

TEIXEIRA, S. M. R. Control of gene expression in Trypanosomatidae. **Brazilian Journal Of Medical And Biological Research**, v. 31, n. 12, p.1503-1516, dez. 1998. FapUNIFESP (SciELO).

TEIXEIRA, S. M. et al. Trypanosomatid comparative genomics: contributions to the study of parasite biology and different parasitic diseases. **Genetics And Molecular Biology**, v. 35, n. 1, p.1-17, 20 jan. 2012. FapUNIFESP (SciELO).

THOMAS, P. D. et al. On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. **Plos Computational Biology**, v. 8, n. 2, p.1002386-1002393, 16 fev. 2012. Public Library of Science (PLoS).

VERLI, H. **Bioinformática:** da Biologia à Flexibilidade Molecular. 1ª Edição. São Paulo: Sociedade Brasileira de Bioquímica e Biologia Molecular - Sbbq, 2014. 292 p.

WAGNER, G. Geração e análise comparativa de sequências genômicas de *Trypanosoma rangeli*. 2006. Dissertação de Mestrado (Biologia Celular e Molecular). Instituto Oswaldo Cruz, Rio de Janeiro, 2006.

WAGNER, G. Análise proteômica de formas tripomastigotas diferenciadas *in vitro* do *Trypanosoma rangeli* e caracterização de antígenos diferenciais ao *Trypanosoma cruzi*. 2012. Tese de Doutorado (Programa de Pós-Graduação em Biotecnologia e Biociências). Universidade Federal de Santa Catarina, Florianópolis.

WAGNER, G. et al. The *Trypanosoma rangeli* trypomastigote surfaceome reveals novel proteins and targets for specific diagnosis. **Journal Of Proteomics**, v. 82, p.52-63, abr. 2013. Elsevier BV.

WAGNER, G. et al. STINGRAY: system for integrated genomic resources and analysis. **Bmc Research Notes**, v. 7, n. 1, p.132-140, 2014. Springer Nature.

WOOD, D. E.; SALZBERG, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. **Genome Biology**, v. 15, n. 3, p.46-57, 2014. Springer Nature.

YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics**, v. 13, n. 5, p.329-342, 18 abr. 2012. Springer Science and Business Media LLC.

ZHANG, Z. et al. PseudoPipe: an automated pseudogene identification pipeline. **Bioinformatics**, v. 22, n. 12, p.1437-1439, 30 mar. 2006. Oxford University Press (OUP).

ZIELEZINSKI, A. et al. Alignment-free sequence comparison: benefits, applications, and tools. **Genome Biology**, v. 18, n. 1, 3 out. 2017. Springer Nature.