CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

INTEGRAÇÃO DE DADOS COMO APOIO A MODELAGEM DE CÉLULA INTEIRA DA BACTÉRIA *PSEUDOMONAS AERUGINOSA* CCBH4851

Ribamar Santos Ferreira Matias

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientadora: Kele Teixeira Belloze

Rio de Janeiro,

Janeiro 2020

INTEGRAÇÃO DE DADOS COMO APOIO A MODELAGEM DE CÉLULA INTEIRA DA BACTÉRIA *PSEUDOMONAS AERUGINOSA* CCBH4851

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Ribamar Santos Ferreira Matias

Banca Examinadora:

Presidente, Professor D.Sc. Kele Teixeira Belloze (CEFET/RJ) (Orientadora)

Professor D.Sc. Eduardo Bezerra da Silva

Professor D.Sc. Fabrício Alves Barbosa da Silva

Rio de Janeiro,

Janeiro 2020

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

M433 Matias, Ribamar Santos Ferreira
Integração de dados como apoio a modelagem de célula inteira da bactéria *Pseudomonas Aeruginosa* CCBH4851/ Ribamar Santos Ferreira Matias.—2020.
101f.: il., color., grafs., tabs., enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2020.
Bibliografia : f. 90-101
Orientadora: Kele Teixeira Belloze

1. Computação. 2. Modelagem Computacional. 3. Bactérias - Identificação. 4. Integração de dados. 5. *Pseudomonas Aeruginosa.* I. Belloze, Kele Teixeira (Orient.). II. Título.

CDD 004

Elaborada pela bibliotecária Teresa Cristina Gaio Mattos – CRB/7 nº 4610

DEDICATÓRIA

Dedico este trabalho a minha mãe, exemplo de perseverança, caráter, honestidade, luta, coragem e determinação; aos meus pais adotivos, que me resgataram para a vida; a minha esposa e meu filho, pelo amor contínuo e apoio de todas as horas, e a todos os meus professores e pessoas que me apoiaram ao longo da vida, que foram luz no meu caminho, permitindo que eu chegasse até aqui.

AGRADECIMENTOS

O presente trabalho foi desenvolvido com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

RESUMO

Integração de Dados como Apoio a Modelagem de Célula Inteira da Bactéria *Pseudomonas aeruginosa* CCBH4851

A análise comparativa de genomas por meio de processos computacionais é uma abordagem de baixo custo e com potencial promissor para apoiar pesquisadores. Tal análise é favorecida ao considerar os diversos dados oriundos de estudos sobre organismos modelo, disponíveis em bancos de dados públicos. Esta abordagem foi utilizada no presente trabalho, para analisar o genoma da cepa *Pseudomonas aeruginosa* CCBH4851. Esta cepa, identificada no Brasil em 2008, está sendo pesquisada pela Fundação Oswaldo Cruz (FIOCRUZ) e parceiros, em função de sua associação a infecções hospitalares, e do seu alto grau de resistência, detectado após testes com diversos antibióticos. Neste sentido, o levantamento de proteínas essenciais, que possam auxiliar no desenvolvimento de novos antibióticos no combate à bactéria, torna-se relevante. Deste modo, o objetivo deste trabalho é construir uma base de dados para ampliar o conhecimento disponível sobre a *Pseudomonas aeruginosa* CCBH4851, a partir de dados provenientes de estudos aprofundados com outros organismos. Esta base de dados reúne informações como anotações por ontologia das proteínas da bactéria, dados sobre homologia e ortologia, e indicadores de similaridade semântica funcional, entre suas proteínas e as de organismos de referência no estudo da espécie *P. aeruginosa*. Como complemento, foi iniciado um processo de aprendizado de máquina, com intuito de inferir quais proteínas da bactéria têm características essenciais, que são o alvo preferencial para ação dos antibióticos. Para reunir este conjunto de informações, foram empregados métodos estritamente computacionais, com o apoio de ferramentas para análise de sequências genômicas, como Blast2GO, InterProScan, GOGO, Blastp e Orthofinder, referenciando conjuntos de proteínas provenientes de bancos de dados genômicos públicos, como Uniprot, OGEE, Interpro e KEGG. O processo de aprendizagem de máquina consistiu na execução de uma rede neural LSTM. Embora sejam menos precisos que as análises por curadoria manual, os métodos computacionais evoluem continuamente, e novas tecnologias e ferramentas para bioinformática são frequentemente disponibilizadas. Estes recursos têm potencial promissor para auxiliar os pesquisadores nas tarefas de conhecimento dos genomas e tomada de decisão. Na base de dados criada, estão disponíveis as anotações pela ontologia Gene Ontology, de aproximadamente 60% do total de proteínas, indicadores de similaridade semântica, assim como o conjunto de proteínas ortólogas da cepa *Pseudomonas aeruginosa* CCBH4851, obtidos através de processos comparativos com proteomas de referência. Por fim, o projeto sugere um fluxo de atividades que pode ser aplicado como abordagem inicial genérica nos estudos de novos genomas, que pode ser aprimorado e estendido por trabalhos futuros.

Palavras-chave: *Pseudomonas aeruginosa* CCBH4851; anotação funcional de proteínas; Gene Ontology

ABSTRACT

Data Integration as Support for Whole Cell Modeling of *Pseudomonas aeruginosa* CCBH4851 Bacteria

Comparative analysis of genomes through computational processes is a low cost approach with promising potential to support researchers. Such analysis is favored by considering the various data from studies on model organisms available in public databases. This approach was used in the present work to analyze the genome of the *Pseudomonas aeruginosa* strain CCBH4851. This strain, identified in Brazil in 2008, is being researched by FIOCRUZ and partners, due to its association with nosocomial infections, and its high degree of resistance, detected after testing with various antibiotics. In this sense, the lifting of essential proteins that may help in the development of new antibiotics in the fight against bacteria becomes relevant. Thus, the objective of this work is to build a database to expand the available knowledge on the *Pseudomonas aeruginosa* CCBH4851, based on data from in-depth studies with other organisms. This database gathers information such as bacterial protein ontology annotations, homology and orthology data, and indicators of functional semantic similarity between their proteins and those of reference organisms in the study of the species *P. aeruginosa*. In addition, a machine learning process was designed to infer which bacteria proteins have essential characteristics, which are the preferred target for antibiotic action. To gather this set of information, strictly computational methods were employed, supported by tools for analysis of genomic sequences, such as Blast2GO, InterProScan, GOGO, Blastp and Orthofinder, referencing sets of proteins from public genomic databases, such as Uniprot, OGEE, Interpro and KEGG. The machine learning process consisted of the execution of an LSTM neural network. Although less accurate than manual curation analysis, computational methods are continually evolving, and new technologies and tools for bioinformatics are often available. These resources have promising potential to assist researchers in genome knowledge and decision making tasks. The Gene Ontology ontology annotations of approximately 60 % of the total proteins, indicators of semantic similarity, as well as the set of orthologous proteins of the *Pseudomonas aeruginosa* CCBH4851 strain, are available in the database created. comparative processes with reference proteomes. Finally, the project suggests a flow of activities that can be applied as a generic initial approach to new genome studies, which can be enhanced and extended by future works.

Keywords: Pseudomonas aeruginosa; neural networks; functional annotation of proteins; data transformation

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE CÓDIGOS

LISTA DE ABREVIATURAS E SIGLAS

BLAST	Basic Local Alignment Search Tool
CDS	Coding Sequence
CI	Conteúdo De Informações
FIOCRUZ	Fundação Oswaldo Cruz
GO	Gene Ontology
INTERPRO	InterPro Consortium
JSON	JavaScript Object Notation
KEGG	KEGG Pathway Database
LSTM	Long Short-Term Memory
NAR	Nucleic Acids Research
NCBI	National Center For Biotechnology Information
OGEE	Online Gene Essentiality
OMS	Organização Mundial De Saúde
PUBMED	US National Library Of Medicine National Institutes Of Health
RNCS	Redes Neurais Convolucionais
RNN	Rede Neural Recorrente
RNNS	Redes Neurais Recorrentes
SGBD	Sistema De Gerenciamento De Banco De Dados
SQL	Structured Query Language
UNIPROT	Universal Protein Resource
W3C	World Wide Web Consortium

SUMÁRIO

Introdução

Infecções hospitalares são um problema em escala global. A cada ano, segundo a Organização Mundial de Saúde (OMS), 4 em cada 10 pacientes são prejudicados nos cuidados de saúde primários e ambulatoriais e 134 milhões de eventos adversos ocorrem em hospitais de países de baixa e média renda, resultando em 2,6 milhões de mortes [World Health Organization, 2019]. Bactérias como as da espécie *Pseudomonas aeruginosa* estão relacionadas a este cenário.

As infecções causadas por esta bactéria se tornaram uma preocupação real em unidades hospitalares, especialmente em pacientes gravemente enfermos e imuno-comprometidos, tornando-se um dos agentes causadores mais frequentes de infecções hospitalares [Klockgether and Tümmler, 2017]. Pesquisas no âmbito de se alcançar novos antibióticos, capazes de combatê-la, são consideradas prioritárias pela OMS, que a qualificou com o grau mais elevado de atenção em sua mais recente Lista Global para Pesquisa, Descoberta e Desenvolvimento de Novos Antibióticos [Tacconelli et al., 2018].

Uma cepa desta espécie, denominada *P. aeruginosa* CCBH4851, foi detectada no Brasil no ano de 2008, e quando submetida a testes, se mostrou resistente a diversos antibióticos. Desde então, os pesquisadores da FIOCRUZ e parceiros buscam identificar alvos para novos medicamentos, capazes de combater esta bactéria [Fundação Oswaldo Cruz, 2019]. Para melhor compreender seus processos biológicos, os pesquisadores planejam elaborar um modelo computacional, conhecido como modelo de célula inteira, capaz de auxiliar na inferência e predição de dados sobre as relações funcionais entre suas estruturas biológicas [Goldberg et al., 2018].

Neste contexto, o objetivo deste trabalho é construir uma base de dados sobre a cepa *Pseudomonas aeruginosa* CCBH4851, com foco na anotação funcional de suas proteínas, para apoiar as tarefas de criação do modelo de célula inteira, conduzidas pela FIOCRUZ. A construção desta base de dados se apoia em quatro premissas: i) oferecer um recurso integrado de informações, para ampliar o conhecimento sobre esta cepa, incluindo dados de organismos de referência no estudo da espécie *Pseudomonas aeruginosa*; ii) centralizar as informações relevantes sobre as proteínas desta cepa, de modo a facilitar as atividades de pesquisa; iii) buscar novas correlações entre os dados, a

partir das informações das fontes acessadas e resultados obtidos; e iv) apoiar, de forma prática, os estudos sobre um problema global de saúde pública, que afeta diariamente a vida de muitos pacientes hospitalizados, principalmente em países como o Brasil.

O foco deste trabalho é ampliar o conhecimento sobre as proteínas da bactéria *Pseudomonas aeruginosa* CCBH4851. O estudo das proteínas tem caráter vital, pois desempenham papéis necessários em quase todos os processos biológicos dos organismos. Muitas respostas para perguntas importantes, como por que temos câncer, por que envelhecemos ou adoecemos, qual a cura para muitas doenças, estão fortemente relacionadas ao estudo e compreensão das proteínas [Gruber et al., 2008]. As seguintes informações são dados relevantes sobre as proteínas da *Pseudomonas aeruginosa* CCBH4851 e integram a base de dados da bactéria:

- Anotações por ontologia: por meio da ontologia Gene Ontology, que é referência para genes e seus produtos, foram anotadas cerca de 60% das proteínas da *P. aeruginosa* CCBH4851.

- Análise de homologia: cerca de 33% das proteínas são consideradas homólogas aos organismos de referência *Escherichia coli* e *Pseudomonas aeruginosa* PAO1; esta é uma característica relevante para inferência funcional, na análise das proteínas.

- Similaridade Semântica: na sequência comparativa entre as proteínas da *P. aeruginosa* CCBH4851 com as dos organismos *E. coli* e *P. aeruginosa* PAO1, observou-se um conjunto de 8% de proteínas funcionalmente similares, utilizando-se a estrutura hierárquica do grafo da Gene Ontology.

Estas informações auxiliam na elaboração do perfil funcional e estrutural da bactéria, extraídas estritamente por meio de resultados de análises computacionais. No contexto dos estudos biológicos, dados que são obtidos desta forma são considerados menos precisos que os provenientes de curadoria manual. No entanto, tarefas manuais de curadoria demandam tempo e custos de mão de obra especializada significativamente maiores, quando comparados as análises informatizadas.

Os recursos computacionais em bioinformática são promissores, frequentemente aprimorados por pesquisadores em todo o mundo, de baixo custo, e oferecem resultados em curto espaço de tempo. A proposta de aliar estes recursos aos conhecimentos disponíveis em diversos bancos de dados públicos curados, pode fornecer dados relevantes

através de análises comparativas, e oferecer resultados que podem auxiliar nos estudos de microrganismos, como a cepa *P. aeruginosa* CCBH4851.

A metodologia elaborada para este trabalho enfatizou o uso de recursos computacionais, visando agregar o conhecimento adquirido dos estudos de outros organismos, por meio de análises comparativas com uso de ferramentas conceituadas de pesquisas genômicas. Do ponto de vista computacional, a abordagem descrita nos procedimentos da metodologia pode servir como um modelo genérico para conhecimento das informações de um genoma, a qual pode ser aplicada aos estudos de quaisquer organismos.

No contexto da presente metodologia, foram pesquisados organismos de referência no estudo da espécie *Pseudomonas aeruginosa*, anotação de genomas, proteínas semanticamente similares do ponto de vista funcional, assim como proteínas com características homólogas entre os organismos estudados. Por fim, foi elaborado um processo computacional, baseado em aprendizado de máquina, que buscou predizer quais proteínas da cepa *P. aeruginosa* CCBH4851 poderiam ter características essenciais. A importância destas proteínas se deve a sua participação nos processos vitais para a sobrevivência do organismo, que são os alvos preferenciais para desenvolvimento de antibióticos.

Esta dissertação está organizada em mais cinco capítulos. O capıtulo 1 fornece o conteúdo necessário para a compreensão dos assuntos abordados neste trabalho. O capıtulo 2 detalha os trabalhos relacionados ao presente trabalho, e a metodologia adotada está especificada no capítulo 3. O capítulo 4 descreve os resultados obtidos, e o capítulo 5 traz as considerações finais sobre os resultados obtidos, bem como limitações encontradas, além de apresentar os cenários futuros e possıveis contribuições.

1- Fundamentação Teórica

Este capítulo apresenta as principais informações necessárias à compreensão deste trabalho. Os conceitos abordados nas seções seguintes abrangem conhecimentos sobre biologia e processos computacionais. Deste modo, para melhor organização e compreensão dos temas, foram elaboradas as seções 1.1 e 1.2, agrupando os pontos específicos apresentados.

1.1- Conceitos biológicos

1.1.1 Bactérias

Bactérias são microrganismos unicelulares de grande interesse médico e científico. Encontram-se em praticamente todos os lugares, são essenciais aos ecossistemas existentes, e consequentemente, aos seres humanos. O corpo humano é repleto de bactérias, e estima-se que sua proporção ocorra na mesma ordem que a das células humanas [Sender et al., 2016].

O primeiro projeto de sequenciamento genômico de uma bactéria foi finalizado no ano de 1995. Segundo [Land et al., 2015], um dos fatores de maior impacto nas pesquisas neste tipo de projeto era o custo, que foi reduzido de forma significativa desde então, permitindo o aumento quantitativo destes projetos nos anos seguintes, bem como o volume de dados e o conhecimento disponível obtido por estas pesquisas, como ilustrado na figura 1.

A maior parte das bactérias é inofensiva, e de certo modo, útil. No entanto, um pequeno número de espécies, conhecido como agentes patógenos, pode causar doenças aos seres humanos [Alberts et al., 2014]. Este grupo de bactérias atrai os interesses médico e científico, em função de diversos problemas que atingem as populações, em escala global, tais como as infecções hospitalares.

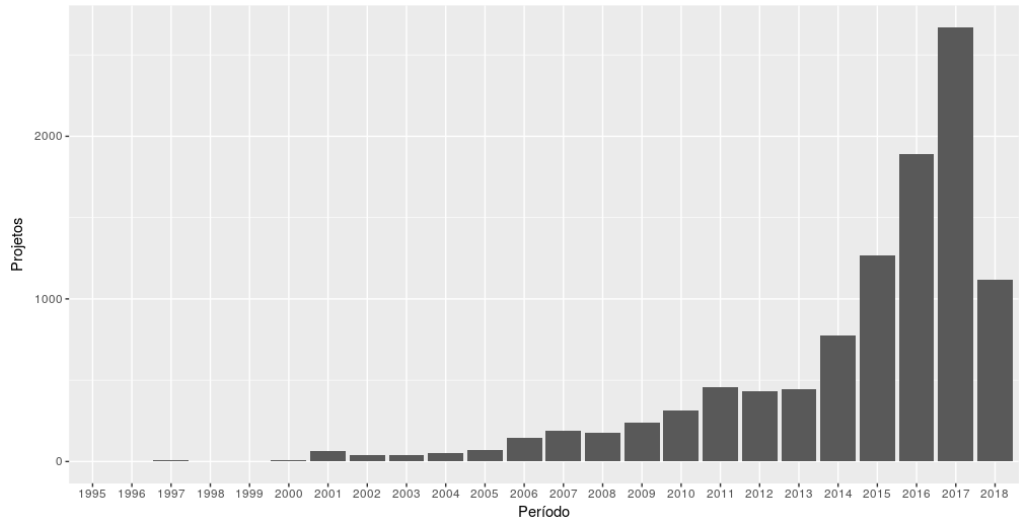


Figura 1 – Projetos de sequenciamento de genomas de bactérias, submetidos ao NCBI. Observa-se mais de 2000 projetos no ano de 2017. [NCBI National Center for Biotechnology Information, 2018b]

Os patógenos podem ser classificados como:

- Obrigatórios: somente podem se reproduzir dentro de células do corpo humano;
- Facultativos: podem se reproduzir em ambientes como a água ou o solo, e somente causam doenças se encontrarem um hospedeiro suscetível;
- Oportunistas: podem causar doenças em um hospedeiro ferido ou imunocomprometido.

As bactérias não são apenas consideradas o berço da vida, mas, como revelado pela história e séculos de interesse científico, são os organismos vivos que mais afetam os humanos [Venkova et al., 2018].

1.1.2 *Pseudomonas aeruginosa*

Bactérias da espécie *Pseudomonas aeruginosa* são patógenos oportunistas, e estão associados a um amplo espectro de infecções humanas, variando de infecções superficiais a sepse fulminante. Em pacientes imunocomprometidos, causam um nível significativo de morbimortalidade. A gravidade das infecções por esta espécie ocorre

em função de fatores predisponentes do hospedeiro, mas também devido a grande variedade de fatores de virulência, assim como a acentuada resistência à maioria dos antimicrobianos usados no uso clínico [da Silva et al., 2018].

A pneumonia hospitalar, infecções da corrente sanguínea e do trato urinário, principalmente em pacientes com queimaduras graves, AIDS, câncer de pulmão, doença pulmonar obstrutiva crônica, bronquiectasias e fibrose cística, são exemplos de infecções causadas por esta espécie de bactérias. A figura 2 ilustra uma amostra da espécie:

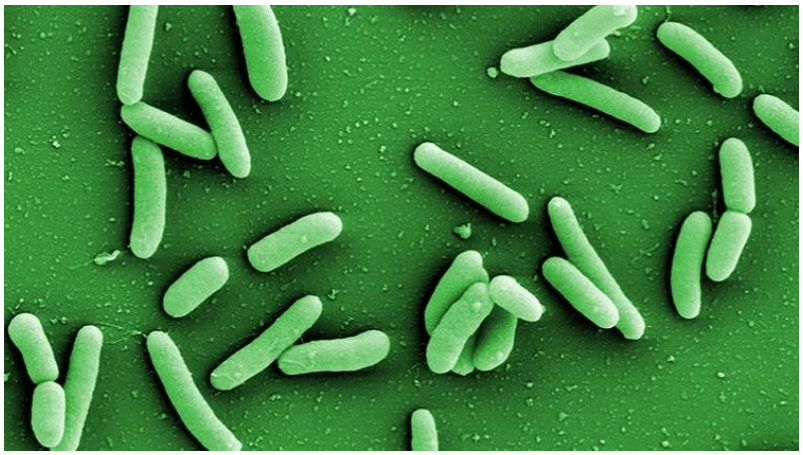


Figura 2 – *Pseudomonas aeruginosa* são extremamente robustas e resistentes e podem residir em praticamente qualquer habitat [DZIF German Center for Infection Research, 2019].

Esta espécie de bactérias é praticamente onipresente, e possui grande versatilidade metabólica e fisiológica, o que lhe permite habitar ambientes terrestres ou aquáticos. São consideradas como uma das três principais causas de infecções humanas oportunistas, o que as consolida como um grave risco à saúde pública. O tratamento para as infecções que estas bactérias causam, em pacientes imunocomprometidos, é particularmente desafiador devido à ampla resistência antimicrobiana intrínseca destas bactérias, fato que dificulta e reduz as opções de tratamentos, dada a disseminação dessa resistência [Vallet-Gely and Boccard, 2013].

Em sua mais recente Lista Global para Pesquisa, Descoberta e Desenvolvimento de Novos Antibióticos, a Organização Mundial de Saúde (OMS) qualificou a espécie *Pseudomonas aeruginosa* com o grau mais elevado de prioridade, para pesquisas que auxiliem a produção de novos antibióticos [Tacconelli et al., 2018].

No contexto do presente trabalho, para fins de análises comparativas com as sequências de proteínas da *P. aeruginosa* CCBH4851, foram utilizadas as informações

de duas bactérias, consideradas referência para estudos da espécie *Pseudomonas aeruginosa*:

- *Pseudomonas aeruginosa* PAO1: foi isolada em 1954 na cidade de Melbourne, na Austrália [Klockgether et al., 2009], e é referência para estudo da espécie *P. aeruginosa*. Esta cepa é o foco principal do banco de dados do Pseudomonas Genome Database, que colabora com um painel internacional de pesquisadores especializados, para fornecer atualizações de alta qualidade com foco nas anotações do seu genoma, e prover dados para análises [Winsor et al., 2015].
- *Escherichia coli*: esta bactéria é o ser vivo mais estudado de todos os tempos, e é o organismo preferido para a investigação da base de mecanismos de genética molecular. A maioria dos nossos conceitos atuais de biologia molecular, incluindo a compreensão sobre a replicação do DNA, sobre o código genético, bem como a expressão gênica e a síntese proteica, derivam de estudos desta bactéria [Cooper and Hausman, 2007]

1.1.3 Genomas

Segundo Goldman and Landweber [2016], o genoma pode ser descrito como um repositório de informações de um organismo. O genoma de todos os organismos vivos consiste em DNA (do inglês *Deoxyribonucleic Acid*), um polímero químico de duas cadeias. Cada cadeia de DNA é composta por quatro unidades diferentes, chamadas nucleotídeos, que estão ligadas de ponta a ponta para formar uma cadeia longa. Estes quatro nucleotídeos são simbolizados como A, G, C e T, que representam as quatro bases, adenina, guanina, citosina e timina - que são partes dos nucleotídeos.

Um passo importante na análise das informações de um genoma é decifrar o potencial de codificação completo ou a CDS (*Coding Sequence*) de proteínas de cada gene. A sigla CDS se refere uma região codificante do gene, correspondente a uma sequência de aminoácidos de uma proteína [Furuno, 2003]. A extração do conjunto de proteínas de um genoma, conhecido como proteoma, é feita com base nas suas sequências codificantes.

1.1.4 Proteínas e proteoma

Proteínas desempenham papéis necessários em quase todos os processos biológicos. Um dos principais objetivos da bioquímica é determinar como as sequências genômicas especificam as conformações, e, consequentemente, as funções das proteínas [Berg et al., 2010].

Processos vitais como o metabolismo, a replicação de DNA, a comunicação célula a célula, a sinalização intracelular, defesa e imunidade, são exemplos de sua importância para os seres vivos. Reações bioquímicas de respiração celular, transporte de oxigênio e gás carbônico, a absorção de alimentos, o uso e armazenamento de energia, as reações fisiológicas ao calor ou frio, são basicamente realizadas por uma proteína ou complexo proteico [Gruber et al., 2008].

De acordo com Garrels [2001], como a maioria das funções enzimáticas celulares, dos reguladores, transdutores de sinal e componentes estruturais são compostos de proteínas, pode-se inferir que as proteínas expressas por uma célula podem fornecer pistas importantes para a função, organização e capacidade de resposta desta célula. Além disso, ao definir a variação entre diferentes células, e entre células expostas a diferentes estímulos, pode-se obter a compreensão sobre assuntos como:

- Adaptação celular a sinais ambientais;
- Mecanismos de diferenciação celular e desenvolvimento organizacional;
- Aspectos celulares dos processos patológicos;
- Respostas celulares ao envelhecimento;
- Diferença entre indivíduos dentro de uma espécie, a base molecular de nossa individualidade em fisiologia, suscetibilidade a doenças e resposta a exposições terapêuticas e ambientais.

A totalidade das proteínas expressas por um genoma, em um dado momento do tempo, sob condições fisiológicas específicas, é definida como proteoma. Num organismo, as células contêm o mesmo genoma; no entanto, elas expressam proteínas diferentes, em resposta a um microambiente específico [Pando-Robles et al., 2009]. O proteoma

pode ser visto como o elo central entre o genoma e a célula: é, por um lado, o ponto mais alto da expressão do genoma e, por outro, o ponto de partida para as atividades bioquímicas que constituem a vida celular [Brown, 2002].

As proteínas são os produtos dos genes e são o material vital e as unidades funcionais dos organismos vivos. Proteínas essenciais são aquelas que atuam em funções biológicas indispensáveis para que os organismos cresçam e se multipliquem normalmente. Assim, a identificação precisa de proteínas essenciais contribui de maneira importante para a compreensão dos principais processos biológicos de um organismo em nível molecular, o que é benéfico tanto para orientar o diagnóstico de doenças, quanto para identificar novos alvos para a criação de medicamentos [Lei and Yang, 2018].

1.1.5 Homologia e ortologia

Homologia é a relação de ancestralidade entre duas ou mais entidades (e.g. genes ou proteínas), ou seja, significa dizer que as mesmas compartilham um ancestral comum [Koonin, 2005]. A identificação de relações de homologia entre sequências genômicas é fundamental para todos os aspectos da pesquisa biológica. As inferências obtidas pelos métodos baseados em homologia apoiam a compreensão da evolução e diversidade da vida. Além disso, elas também fornecem uma estrutura coerente para a extrapolação do conhecimento biológico entre organismos [Emms and Kelly, 2015].

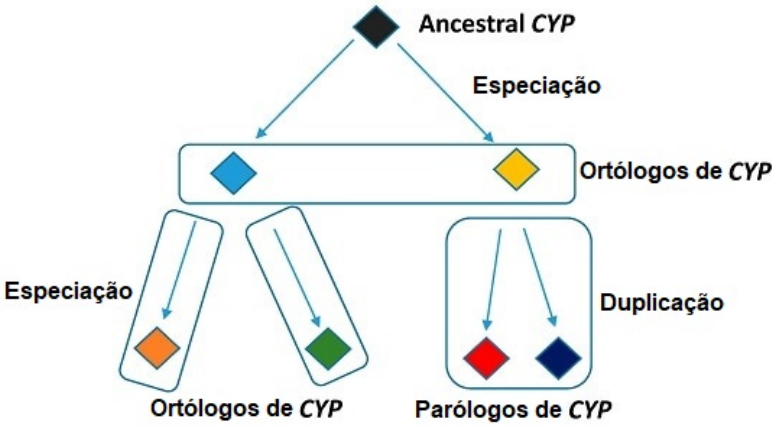


Figura 3 – Conceito de ortologia, figura adaptada de [Brennan, 2019].

Ortologia é um caso especial de homologia, como ilustrado na figura 3. Neste caso, os genes são derivados de um único gene ancestral, no último ancestral comum entre espécies comparadas [Koonin, 2005], que no exemplo é o ancestral CYP.

De acordo com o conceito, proteínas de diferentes organismos que se dividiram por especiação na evolução podem compartilhar a mesma função. Desta forma, havendo uma proteína essencial de um organismo modelo que seja ortóloga a uma proteína de um organismo de estudo, faz sentido dizer, que esta última, possa ter características de essencialidade também. A análise de grupos ortólogos é de grande importância para a biologia computacional, anotação de genomas e inferência filogenética [Nichio et al., 2017].

Informações sobre proteínas ortólogas são outro aspecto importante para a identificação de proteínas essenciais. As proteínas ortólogas são proteínas derivadas de um ancestral comum e geralmente mantêm as mesmas funções ou funções muito semelhantes. Está provado que as propriedades ortólogas estão positivamente correlacionadas com a essencialidade de uma proteína [Qin et al., 2017].

1.2- Conceitos computacionais

1.2.1 Ontologias

Ontologias são vocabulários para representar as definições de um domínio compartilhado, por meio de suas classes, relacionamentos, funções e outros objetos [Gruber, 1993]. Elas são utilizadas para descrever e classificar entidades de interesse, como por exemplo, processos biológicos.

De acordo com o World Wide Web Consortium (W3C), principal órgão de padrões da internet, o conceito de ontologia se aplica a coleções com termos formais, com grau significativo de complexidade [W3C World Wide Web Consortium, 2018]. Esta complexidade é comum na classificação de genes, cujas pesquisas lidam com grandes volumes de informação, e que, consequentemente, dependem do apoio de sistemas computacionais.

Neste sentido, a anotação de proteínas por homologia elaborada neste projeto utilizou a ontologia Gene Ontology (GO). A Gene Ontology (GO) é uma ontologia específica para assuntos relacionados a informações sobre genes e seus produtos [Gene Ontology Consortium, 2019]. Seu objetivo é permitir o acesso das classificações biológicas aos processos computacionais, por meio de um esquema organizado, uniforme e hierárquico de classificações [Ashburner et al., 2000].

O projeto GO teve início em 1998, e representa um esforço colaborativo, com foco em dois aspectos, voltados à integração de dados: prover descritores consistentes para produtos genéticos, e padronizar as classificações para sequências genômicas e suas características. Inicialmente contou com dados de três organismos modelos, e hoje integra diversos bancos de dados, incluindo repositórios com dados de plantas, animais e microrganismos. O banco de dados GO integra os vocabulários e as anotações provenientes de contribuições, oferecendo acesso a estas informações em diversos formatos. Os membros do Consórcio Gene Ontology trabalham de modo contínuo e coletivamente, envolvendo o apoio de especialistas externos quando necessário, para expandir e atualizar seu vocabulário [Consortium, 2004].

A GO é considerada a maior fonte de recursos para catálogo de funções genômicas [du Plessis et al., 2011], e é a base de conhecimento mais abrangente e amplamente utilizada sobre as funções dos genes [The Gene Ontology Consortium, 2018]. Seu banco de dados está estruturado nas categorias processo biológico, componente celular e função molecular, que caracterizam diferentes conceitos biológicos de um organismo:

- Função Molecular: é o conjunto de atividades que o gene desempenha;
- Processo Biológico: é o conjunto de processos que o gene participa;
- Componente Celular: é o conjunto de informações que descreve a localização do gene.

A estrutura da GO pode ser descrita em termos de um grafo, como ilustrado na figura 4. Cada termo GO é um nó e as arestas são os relacionamentos entre os termos. O grafo GO é vagamente hierárquico, com os termos filhos sendo mais especializados que os termos pais, mas diferentemente de uma hierarquia estrita, um termo pode ter mais de um termo pai. Neste exemplo, o modelo pai e filho não é verdadeiro para todos os tipos de relações. Observando o termo do processo biológico, *hexose biosynthetic process*,

identificado pelo código GO:0019319. Verifica-se a existência de dois progenitores, os termos GO:0019318 e GO:0046364, o que indica que é subtipo de ambos [Gene Ontology Consortium, 2019].

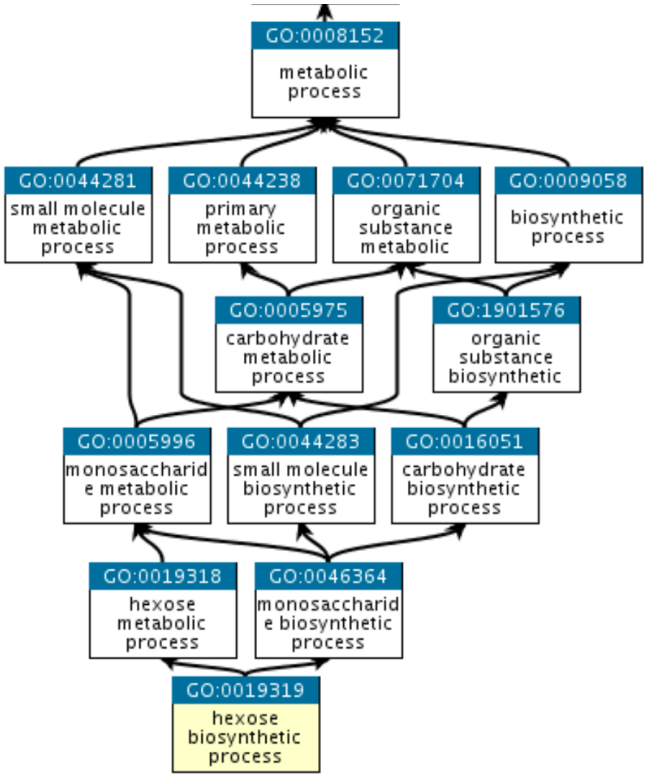


Figura 4 – Exemplo de estrutura de grafo de anotação Gene Ontology [Gene Ontology Consortium, 2019].

1.2.2 Anotação de sequências genômicas

Anotar uma sequência genômica consiste em encontrar e descrever as localizações individuais de cada gene, e outras características, com base no genoma original do micro-organismo em estudo. As anotações dão significado a uma determinada sequência, e facilitam sua compreensão, para pesquisadores visualizarem e compreenderem seu conteúdo [NCBI National Center for Biotechnology Information, 2019].

Anotações abrangentes de recursos proteicos são uma maneira eficaz de construir uma imagem da função da proteína. Tais características podem incluir: genes, os níveis de expressão, a posição de elementos reguladores, locais de ligação, processamento do RNA

e variância individual; e, para proteínas, posição de resíduos funcionais, identificação de modificações pós-traducionais, descrição de resíduos que interagem com DNA, proteína ou ligante, elucidação ou predição dos parceiros do domínio, descrição da unidade biológica global e até dados que descrevem a estrutura dimensional da proteína [Reeves et al., 2009].

Este trabalho utiliza processos de anotação por homologia, com base em ontologias, como detalhado na seção 1.2.1. As anotações baseadas em ontologia servem como descritores dos genes e proteínas, e permitem a localização de informações relevantes destas estruturas por meio de buscas utilizando palavras-chave, nos diversos bancos de dados genômicos disponíveis. Ademais, as anotações baseadas em ontologia apoiam a integração, o compartilhamento dos dados e o trabalho colaborativo.

Anotações podem ser feitas por curadoria manual, ou por processos de predição computacionais. No contexto computacional que é o alvo deste trabalho, existem duas classes principais de métodos para predição de proteínas. Uma é baseada em pesquisas de similaridade de sequência, enquanto a outra observa a estrutura de genes e pesquisas baseadas em sinais, que também é conhecida como descoberta de genes *ab initio* [Wang et al., 2004].

O método de anotação por curadoria manual fornece os conjuntos de dados de maior precisão; no entanto, demanda mais tempo e pode cobrir apenas uma fração dos dados a serem anotados. Alternativamente, a informação pode ser obtida transferindo-se o conhecimento existente sobre uma sequência para a sequência relacionada, considerada homóloga (conceito definido na seção 1.1.5). A precisão desses métodos depende da distância evolutiva; quanto maior a distância, menor a confiança que pode-se ter em prever com precisão um recurso. Por fim, algumas anotações podem ser previstas usando métodos *ab initio*, que usam regras treinadas em anotações anteriores ou as propriedades físico-químicas da molécula para prever a característica [Reeves et al., 2009].

Em geral, ao escolher um método de anotação, é preciso ponderar as demandas frequentemente concorrentes de velocidade e precisão. Curadoria manual ou métodos experimentais têm alta precisão, mas consomem tempo, e provavelmente são mais apropriados para conjuntos de dados pequenos. Os métodos que produzem anotações com maior velocidade e cobertura (por exemplo, transferência por homologia) frequentemente o fazem com menor precisão, mas seu uso pode ser mais indicado quando os conjuntos de dados são grandes [Reeves et al., 2009].

A figura 5 exemplifica dois momentos, antes e após o procedimento de anotação via ontologia GO, das proteínas *PA4851_00055* e *PA4851_00060*. Na parte superior da figura, em branco, constam somente a sequência de proteína e sua descrição, e na parte inferior, em azul, as mesmas sequências, já com suas anotações por ontologia, capturadas pelas ferramentas genômicas.

SeqName	Description	Length	#Hits	e-Value	sim mean	#GO	GO IDs	GO Names
PA4851_00055|lysophosphatidic	acid acyltransferase|COORDINATES: similar to AA sequence:RefSeq:NP_064725.1	257						
PA4851_00060|D-glycero-beta-...	7-phosphatase|COORDINATES: similar to AA sequence:RefSeq:NP_064726.1	178						

SeqName	Description	Length	#Hits	e-Value	sim mean	#GO	GO IDs	GO Names
PA4851_00055|lysophosphatidic	LPAT1_BRANARecName: Full=1-acyl-sn-glycerol-3-phosphate acyltransferase BAT2, chloroplastic; Alt...	257	20	3.25E-22	47.27%	7	P:GO:0046474; P:GO:0048518; P:GO:0048856; F:GO:0016746; C:GO:0009507; C:GO:0012505; C:GO:0031090	P:glycerophospholipid biosynthetic process; P:positive regulation of biological process; P:anatomical structure development; F:transferase activity, transferring acyl groups; C:chloroplast; C:endomembrane system; C:organelle membrane
PA4851_00060|D-glycero-beta-...	GMHBB_PSEAERecName: Full=D-glycero-beta-D-manno-heptose-1,7-bisphosphate 7-phosphatase; A...	178	20	4.13E-128	60.77%	7	P:GO:0009244; P:GO:0016311; P:GO:0097171; F:GO:0000287; F:GO:0008270; F:GO:0034200; C:GO:0005829	P:lipopolysaccharide core region biosynthetic process; P:dephosphorylation; P:ADP-L-glycero-beta-D-manno-heptose biosynthetic process; F:magnesium ion binding; F:zinc ion binding; F:D,D-heptose 1,7-bisphosphate phosphatase activity; C:cytosol

Figura 5 – Exemplo antes e após da anotação por ontologia

No exemplo, ambas as proteínas foram anotadas com sete termos GO. Os prefixos F:, P: e C:, indicam as categorias Função Molecular, Processo Biológico e Componente Celular. Quando anotada, a proteína pode conter um ou mais termos, não necessariamente em todas as categorias GO.

1.2.3 Similaridade semântica

Quando entidades biológicas são descritas usando um esquema comum, como uma ontologia, elas podem ser comparadas por meio de suas anotações. Este tipo de comparação é chamado similaridade semântica, pois avalia o grau de parentesco entre duas entidades pela semelhança no significado de suas anotações [Pesquita et al., 2009]. A figura 6 ilustra um exemplo de similaridade semântica.

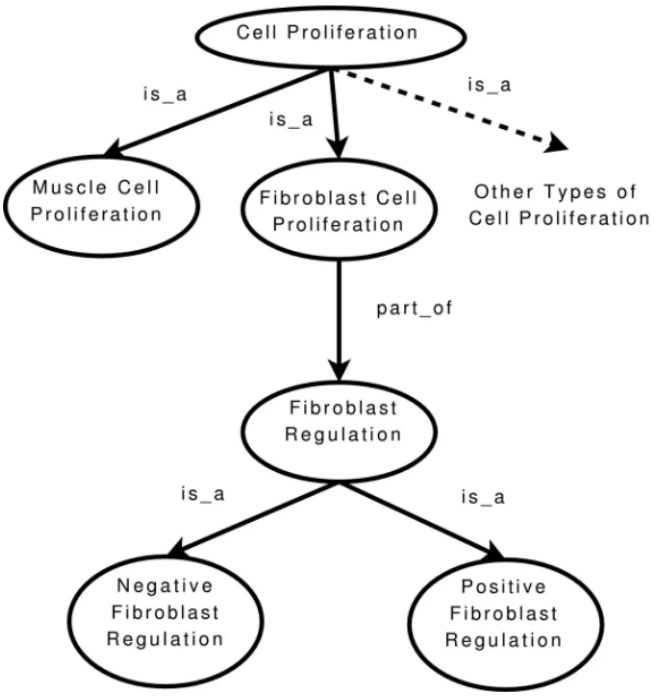


Figura 6 – Exemplo de similaridade semântica [Sheehan et al., 2008].

Em termos de distância do grafo, de acordo com a figura 6, podemos considerar os termos *Muscle Cell Proliferation* e *Fibroblast Cell Proliferation* como sendo mais semelhantes do que o termo anterior com *Fibroblast Regulation*. No entanto, a distância do grafo tem apenas uma fraca correlação com a similaridade dos termos. A semelhança

semântica entre *Positive Fibroblast Cell Regulation* e *Negative Fibroblast Cell Regulation* é muito maior que a similaridade entre *Muscle Cell Proliferation* e *Fibroblast Cell Proliferation*, embora ambos os exemplos tenham distância dois [Sheehan et al., 2008].

Neste sentido, o conhecimento sobre a similaridade semântica entre entidades biológicas, como as proteínas, auxilia no enriquecimento da anotação funcional. A inferência de similaridades semânticas entre os termos da ontologia GO é considerada um componente fundamental na pesquisa em bioinformática funcional, como agrupamento de genes, previsão de funções proteicas e validações de interações proteína-proteína [Zhao and Wang, 2018].

1.2.4 Bancos de dados biológicos

Os bancos de dados desempenham um papel cada vez mais importante na biologia. Eles arquivam, armazenam, mantêm e compartilham informações sobre genes, genomas, dados de expressão, sequências e estruturas de proteínas, metabólitos e reações, interações e vias metabólicas [Zhulin, 2015].

A crescente oferta de bancos de dados públicos biológicos se tornou um importante aliado à pesquisa científica. Em dezembro de 2018, a Nucleic Acids Research (NAR) destacou a existência de 66 novos bancos de dados, de um total de 1613 existentes [Rigden and Fernández, 2018]. Os bancos citados a seguir, forneceram a base de informações necessária para as várias etapas da metodologia:

- InterPro Consortium (InterPro): este banco tem foco na classificação de sequências e famílias de proteínas, para auxiliar nas tarefas de predição de domínios. Agrega informações de outros bancos de dados genômicos, mantidos por centros de pesquisa de referência [Mitchell et al., 2018].

- GO: é um recurso abrangente de conhecimento computacional, sobre as funções de genes e produtos gênicos [Gene Ontology Consortium, 2016]. Fornece vocabulários e classificações estruturados e controlados que abrangem vários domínios da biologia molecular e celular e estão disponíveis gratuitamente para uso da comunidade, na anotação de genes, produtos e sequências de genes [Consortium, 2004].

- KEGG Pathway Database (KEGG): é um recurso integrado para interpretação biológica de sequências e genomas. Funções moleculares de genes e proteínas estão associadas a grupos ortólogos e armazenadas no banco de dados KEGG [Kanehisa et al., 2015].

- Online Gene Essentiality (OGEE): é um banco de dados que fornece dados sobre genes com características essenciais. Genes essenciais participam dos processos vitais para a sobrevivência do organismo. Este banco contém 167.799 genes, testados por essencialidade a partir de 48 espécies, acrescido por 91.000 genes e 24 espécies, respectivamente, a partir de sua última atualização [Chen et al., 2016].

- Universal Protein Resource (UniProt): A base de conhecimento UniProt é um grande repositório de sequências de proteínas e anotações detalhadas. Este banco de dados contém mais de 60 milhões de sequências, das quais mais de meio milhão foram selecionadas por especialistas, que revisam criticamente os dados experimentais e previstos para cada proteína [The UniProt Consortium, 2018].

1.2.5 Ferramentas genômicas

Os softwares de biologia computacional são amplamente difundidos e apoiam a produção de algumas das publicações mais citadas no *corpus* científico. São ferramentas que implementam métodos para alinhamento de sequências e inferência de homologia, análise filogenética, análise estatística de padrões em biomedicina, análise de estrutura biomolecular, processos de visualização e coleta de dados [Gardner et al., 2016].

O repositório OMICtools é um exemplo de recurso para busca de ferramentas genômicas [Henry et al., 2014]. Uma pesquisa utilizando a string de busca *protein annotation*, em novembro de 2019, retornou 17275 resultados. Filtrando-os pela categoria *Gene set enrichment analysis*, que é relacionada ao enriquecimento de informações sobre um genoma, foram obtidos 157 resultados, dos quais a maioria é composta por ferramentas de uso livre, que é uma importante característica. As plataformas de uso livre são preferidas aos produtos proprietários devido ao custo e à personalização. Comparados aos produtos proprietários, os de uso livre não exigem taxas de licença, podem ser

adaptados para atender às necessidades do projeto, contam com o apoio da comunidade de desenvolvimento, e contribuem para ampliar o conhecimento [Kanter et al., 2012].

As tarefas da metodologia empregaram as seguintes ferramentas de bioinformática:

- BioPython [Cock et al., 2009]: é um conjunto de ferramentas disponíveis gratuitamente para computação biológica, escritas em Python por uma equipe internacional de desenvolvedores. O objetivo do Biopython é facilitar ao máximo o uso do Python para bioinformática, criando módulos e classes reutilizáveis de alta qualidade.

- Blast2GO [Gotz et al., 2008]: é uma plataforma de bioinformática para anotação funcional de alta qualidade e análise de conjuntos de dados genômicos. Permite analisar e visualizar genomas recém-sequenciados, combinando metodologias de ponta, recursos padrão e algoritmos.

- InterProScan [Jones et al., 2014]: é uma ferramenta usada para análise de sequência de proteínas e nucleotídeos. A base de dados de referência para pesquisa de sequências genômicas desta ferramenta, é o banco de dados InterPro, que é um conglomerado de bancos de dados, para classificação e predição de sequências de proteínas.

- Orthofinder [Emms and Kelly, 2015]: esta ferramenta analisa sequências de proteínas, processando estatísticas genômicas comparativas com espécies diferentes, para inferir se estas sequências provém de um ancestral comum.

- GOGO [Zhao and Wang, 2018]: analisa a similaridade semântica entre conjuntos de proteínas. Indicadores de similaridade semântica permitem inferir se uma proteína é similar a outra, do ponto de vista funcional.

- Bibliotecas para aprendizado de máquina: foram utilizadas conjuntos de rotinas para aprendizado de máquina, scikit-learn [Pedregosa et al., 2012] e keras [Chollet et al., 2015], a partir da plataforma Google Colab.

- Banco de Dados MySQL: é um banco de dados padrão relacional, apontado como o segundo mais utilizado da atualidade (conforme seção 1.2.6). As informações obtidas ao longo das várias fases de processamento, indicadas na metodologia, foram armazenadas neste banco de dados.

A escolha para as ferramentas utilizadas observou critérios como referências e citações em publicações de pesquisas. Pesquisando no repositório US National Library of Medicine National Institutes of Health (PubMed) as palavras-chave ”Blast2GO”e ”InterProScan”, em Fevereiro de 2020, foram encontrados como resultados, 172 e 143 ocorrências, respectivamente, indicando citações de uso destas ferramentas em projetos. O repositório PubMed é um recurso de referência para pesquisas biológicas, que compreende mais de 30 milhões de citações para literatura biomédica do MEDLINE, periódicos de ciências da vida e livros on-line [National Center for Biotechnology Information (US), 2019].

Tarefas de anotação de sequências genômicas são exemplos de atividades que demandam apoio de ferramentas, em função do extenso processamento envolvido. As ferramentas desempenham papel cada vez mais relevante em pesquisas biológicas, seja em função do crescente volume de dados que manipulam, bem como da necessidade de processar tais conjuntos de dados de forma cada vez mais ágil, com desempenho satisfatório. Segundo Kearse et al. [2012], as duas principais funções da bioinformática são a organização e análise dos dados biológicos, por meio de recursos computacionais.

1.2.6 Criação de bancos de dados

Bancos de dados são repositórios de informações, utilizados pela maioria dos sistemas de informação atuais. A criação de um banco de dados biológico não é, em princípio, um processo diferente da criação de um banco de dados para uma empresa comercial, uma instituição financeira, ou agência governamental [Birney, 2004]. Assim como em todos estes cenários, é necessário compreender as informações que farão parte do banco de dados, suas regras de armazenamento, e as relações entre as informações armazenadas, para que o conjunto se traduza em uma estrutura de fácil uso para aqueles que irão acessar seu conteúdo, de modo que possam fazê-lo de forma simples e direta, e extrair rapidamente as informações de que necessitam.

Um Sistema de Gerenciamento de Banco de Dados (SGBD) é uma coleção de dados inter-relacionados e um conjunto de programas para acessar esses dados [Silberschatz et al., 2001]. O aprimoramento destes programas cresceu significativamente,

impulsionado pela alta demanda de uso dos sistemas de informação, que atualmente armazenam dados estruturados e não estruturados, e precisam prover acesso a consultas e armazenamento de grandes volumes de dados. Tecnologias para redução do tempo de acesso aos dados, assim como para otimizar as formas de armazenamento, são fatores pesquisados e constantemente aprimorados, para prover melhor desempenho aos sistemas de informação que utilizam bancos de dados.

Existem atualmente diversos tipos de SGBDs, de uso comercial e livre, com os quais é possível criar estruturas para atender desde projetos simples a aqueles com alto grau de complexidade. O principal produto deste trabalho é um banco de dados, criado com base nas informações da cepa *P. aeruginosa* CCBH4851, obtidas por meio dos resultados produzidos pelas ferramentas e análise dos dados, nas tarefas de análise de sequncias genômicas, empregadas nas etapas da metodologia proposta. Os principais critérios de escolha para o SGBD orientado a criação do banco de dados deste projeto foram:

- Facilidade de uso: como o público alvo deste trabalho são pesquisadores, este ponto teve relevância na escolha. Atualmente, existem diversas tecnologias para armazenamento de dados, que exploram modelos consolidados como o relacional, presente na maioria das aplicações em uso na atualidade, e novas tendências, como os modelos NoSQL, que exploram cenários onde o modelo relacional necessita de melhorias. Neste projeto, a proposta foi elaborar um modelo inicial, visando acesso simplificado aos dados, e que ao mesmo tempo pudesse ser facilmente estendido a novos modelos. Neste contexto, o banco de dados MySQL foi o mais indicado, em função do seu uso em larga escala. Este banco de dados possui instaladores para as plataformas mais utilizadas do mercado, e pode ser executado mesmo em computadores com baixo poder de processamento.

- Versões de uso livre: desenvolvedores tradicionais de gerenciadores de bancos de dados, como Oracle, Microsoft e IBM, oferecem versões de bancos de dados com recursos limitados, restritas a períodos de testes. Versões de uso livre não contém tais restrições, e podem favorecer a troca de dados entre pesquisadores. Sobre o banco de dados MySQL, não há limites quanto ao número de bancos gerenciados, tabelas, e nem ao volume de dados inserido [MySQL, 2020].

- Disponibilidade de ferramentas: ferramentas para extração e consultas são um

diferencial para bancos de dados. Neste contexto, existem diversas ferramentas de uso livre, que se conectam com facilidade ao banco de dados MySQL e permitem a elaboração de consultas e manipulação de dados.

Fontes como Solid IT [2019] e a revista DeveloperWeek [Scalegrid.io, 2020], indicam que o MySQL é um dos bancos de dados mais utilizados e populares da atualidade, como indicado na figura 7.

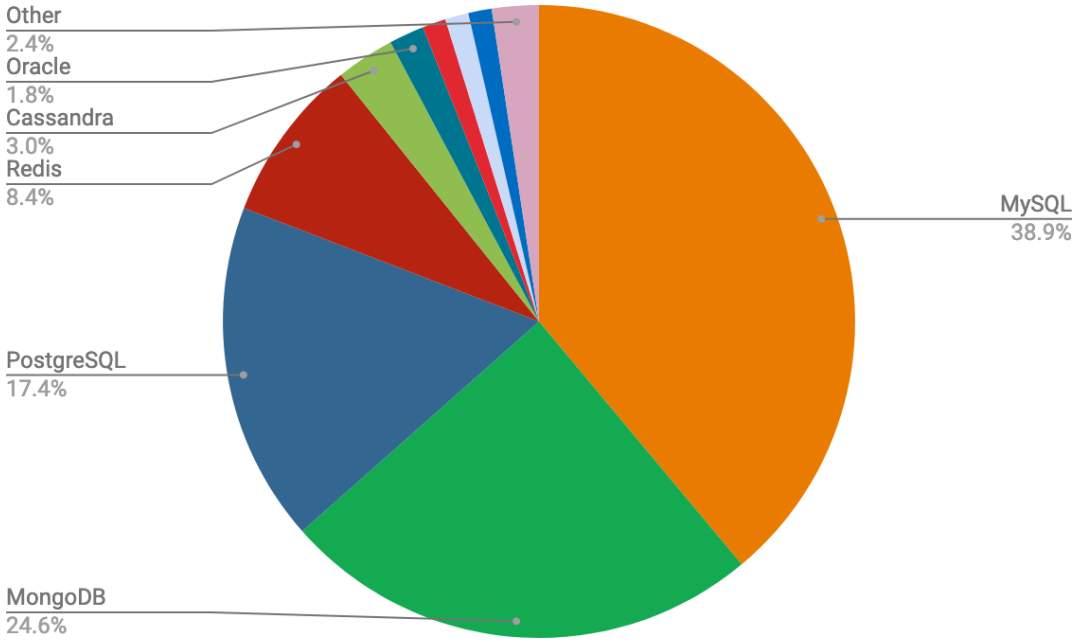


Figura 7 – Bancos de dados mais populares de 2019

Os indicadores divulgados pela revista DeveloperWeek são reforçados pelos resultados da pesquisa anual promovida pelo site Stack Overflow, que é uma das maiores comunidades técnicas para troca de informações entre desenvolvedores. Os resultados da pesquisa de 2019, descritos na figura 8, apontam o MySQL como o SGBD mais utilizado nos últimos dois anos, e consequentemente, uma boa opção para armazenamento e troca de dados [Stack Overflow, 2020].

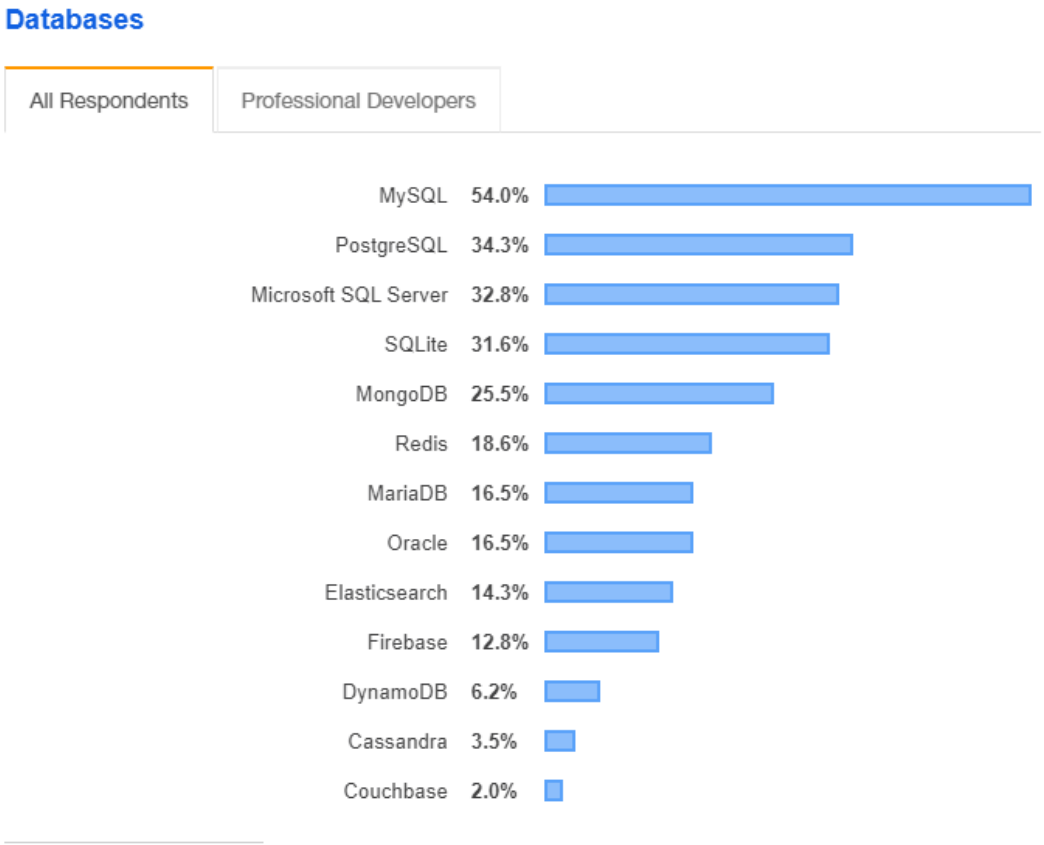


Figura 8 – Ranking dos bancos de dados mais utilizados em 2019[Stack Overflow, 2020]

O SGBD MySQL pode ser acoplado de modo simplificado a outros sistemas, especialmente a projetos voltados para uso via internet. Este banco de dados está disponível nas versões comercial e comunitária,esta última empregada neste projeto.

Além das informações provenientes dos resultados de processamento das ferramentas e análise dos dados, também fazem parte do banco de dados os conjuntos de proteínas extraídos de outros recursos, como Uniprot, que forneceu os conjuntos de proteínas das bactérias *E. coli* e *P. aeruginosa* PAO1; e OGEE, provendo as proteínas essenciais para treinamento e testes da rede neural.

1.2.7 Redes neurais

Um processo computacional, baseado em aprendizado de máquina, foi elaborado com o objetivo de predizer quais proteínas da cepa *P. aeruginosa* CCBH4851 podem ter

características essenciais. Este processo empregou uma rede neural recorrente Long Short-Term Memory (LSTM). Uma visão geral sobre redes neurais é apresentada na subseção 1.2.7.1, enquanto que os conceitos sobre redes recorrentes e arquiteturas LSTM são abordados nas subseções 1.2.7.2 e 1.2.7.3. O processo de aprendizado de uma rede neural é descrito na subseção 1.2.7.4, e a subseção 1.2.7.5 indica as métricas mais utilizadas para avaliação de desempenho dos resultados obtidos via processamento de uma rede neural.

1.2.7.1 Visão geral

Tecnologias emergentes, como aprendizado de máquina e aprendizagem profunda, que abrangem o campo de redes neurais, estão alcançando sucesso em diversos cenários. No campo da bioinformática, a popularização e expansão recentes das ferramentas de sequenciamento facilitaram investigações genômicas em larga escala e comparações entre espécies. Em particular, os algoritmos de aprendizado de máquina estão aprimorando tanto as análises comparativas entre genes, quanto as predições de essencialidade, explorando recursos que diferenciam genes essenciais de não essenciais [Campos et al., 2019]. As predições baseadas em aprendizado de máquina, podem fornecer informações valiosas sobre as funções das proteínas, assim como a ocorrência de doenças e o tratamento para as mesmas [Li et al., 2018].

Uma rede neural é uma técnica flexível e relevante de aprendizado de máquina, que busca imitar o cérebro humano no processamento de sinais de entrada, transformando-os em sinais de saída [Zhang, 2016]. Estas redes são conjuntos interconectados de elementos, unidades ou nós de processamento, cuja funcionalidade é baseada em neurônios cerebrais. A capacidade de processamento da rede é armazenada nos pontos fortes da conexão entre unidades, ou pesos, obtidos por um processo de adaptação, ou aprendizado, de um conjunto de padrões de treinamento [Gurney, 2014]. A figura 9 ilustra os modelos biológico e matemático de um neurônio.

No referido exemplo, o neurônio artificial recebe informações por meio dos pontos de entrada do seu corpo celular, e executa um processamento interno com estas informações. Após o processamento a que foi destinado, o resultado de sua execução

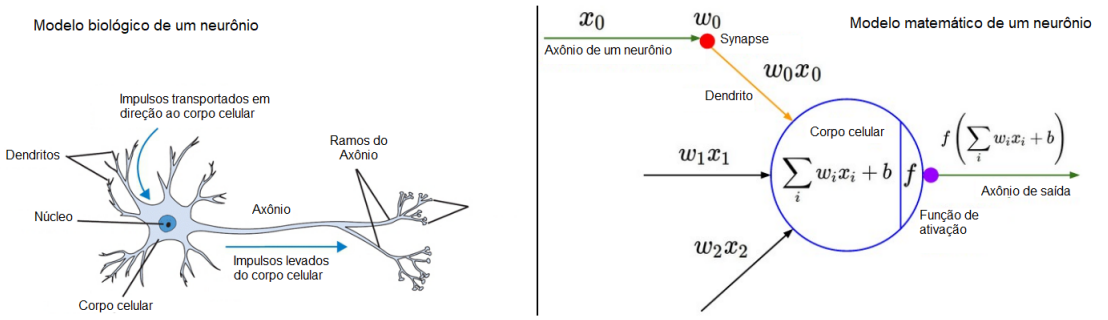


Figura 9 – Modelos biológico e matemático do neurônio, adaptado de [Stanford University, 2019].

é testado por uma função de ativação, que pode transmitir novos resultados aos nós seguintes, ou encerrar a execução da rede como um todo. A figura 10 exemplifica um processamento similar por rede neural.

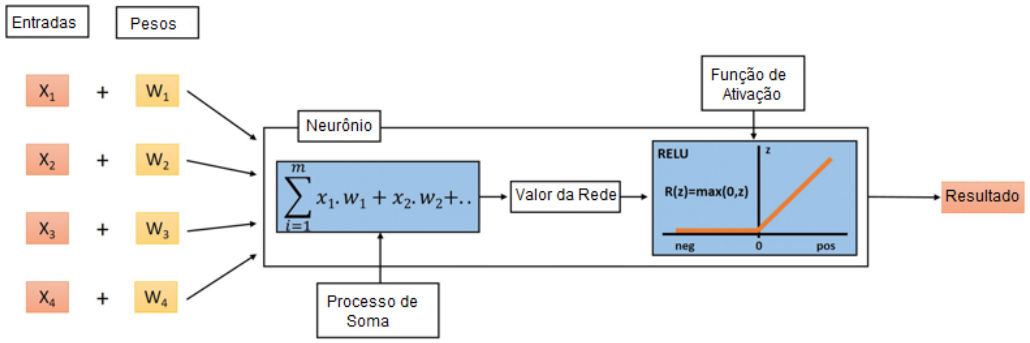


Figura 10 – Exemplo de processamento via rede neural, adaptado de [Georgevici and Terblanche, 2019]

Este exemplo aborda a soma e ativação dentro de um único neurônio computacional. Dentro de cada neurônio, os valores de entrada são combinados com um peso, e somados antes de serem passados para uma função de ativação. A saída do neurônio depende do tipo de função de ativação usada. Na figura, um valor somado acima de zero produzirá uma saída proporcional ao valor de entrada, enquanto uma entrada de ativação igual ou inferior a zero gera um valor de saída para esse neurônio de zero. O aprendizado é obtido atualizando-se repetidamente os valores de peso, durante a retropropagação até encontrar a combinação de valores que melhor mapeia as entradas, para a principal saída de interesse [Georgevici and Terblanche, 2019].

1.2.7.2 Redes neurais recorrentes

Existem diferentes arquiteturas de redes neurais, e cada uma oferece vantagens e desvantagens, em função do cenário em que são aplicadas: Redes Neurais Convolucionais (RNCs), Redes Neurais Recorrentes (RNNs), *feedforward* e *autoencoder* [SAS, 2019]. A figura 11 ilustra uma arquitetura básica de uma rede neural recorrente.

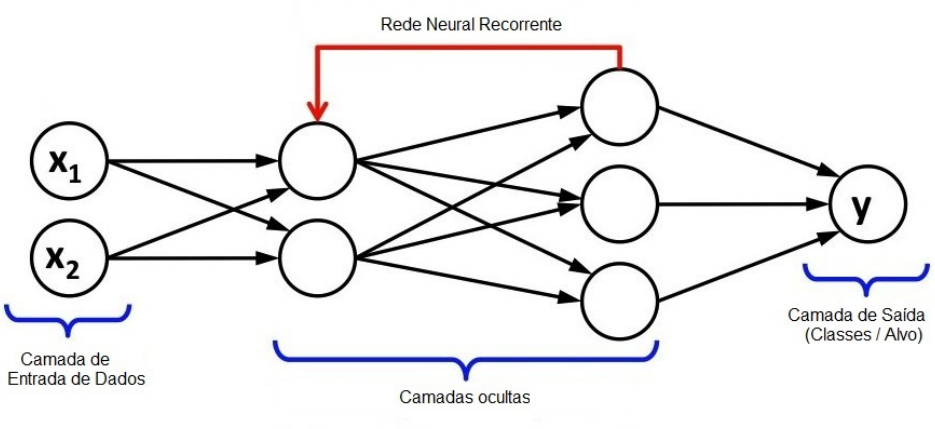


Figura 11 – Exemplo de rede neural recorrente [Chatterjee, 2019].

A ideia por trás das RNNs é fazer uso de informações sequenciais. Em uma rede neural tradicional, assume-se que todas as entradas (e saídas) são independentes uma da outra. Supondo que um processo queira prever a próxima palavra em uma frase, é importante saber quais palavras vieram antes dela. Redes neurais RNNs são chamadas de recorrentes porque executam a mesma tarefa para todos os elementos de uma sequência, com a saída sendo dependente dos cálculos anteriores. Uma forma simplificada de compreender o funcionamento destas redes é pensar que elas possuem uma memória que captura informações sobre o que foi calculado até o momento [Britz, 2016].

Redes neurais recorrentes sofrem de memória de curto prazo. Se uma sequência for longa o suficiente, a rede encontra dificuldade em transportar informações das etapas anteriores para as posteriores. Um dos padrões criados para solucionar este problema é o LSTM. Este padrão de rede neural conta com apoio de estruturas chamadas portões, que regulam o fluxo de informações durante as execuções. Estas estruturas podem aprender quais dados são necessários, bem como os que podem ser descartados, durante o processo de aprendizagem [Mohamed Elfil, 2019].

1.2.7.3 Redes LSTM

Este trabalho faz uso exclusivamente de uma arquitetura de redes neurais recorrentes, conhecida como LSTM. Uma rede neural LSTM é um tipo de Rede Neural Recorrente (RNN) que incorpora um mecanismo altamente eficaz para determinar quais elementos do estado codificado devem ser transmitidos para a próxima célula a cada momento e quais usar para prever a variável alvo. Embora a RNN básica seja uma rede útil para lidar com informações sequenciais, a rede LSTM possui certas vantagens distintas sobre as unidades RNN simples. Redes neurais RNN simples tendem a exibir degradação ao aprender sequências de entrada longas, em que a rede não tem a oportunidade de redefinir seu estado interno. Uma rede RNN simples também pode falhar na previsão de resultados quando as informações mais críticas na sequência estão a muitos passos de distância da janela de tempo prevista. No entanto, as redes LSTM modernas contêm portões de entrada, saída e esquecimento. Os portões de esquecimento permitem que a rede aprenda sequências longas, lide com dependências de longo alcance e possa convergir em soluções relevantes no contexto avaliado [Kaji et al., 2019].

As redes LSTM são redes recorrentes onde cada neurônio é substituído por uma unidade de memória. A unidade de memória contém um neurônio real com uma auto-conexão recorrente. As ativações desses neurônios nas unidades de memória são chamadas de estado celular da rede LSTM [Towards Data Science, 2020]. A figura 12 ilustra a arquitetura de uma unidade de memória LSTM.

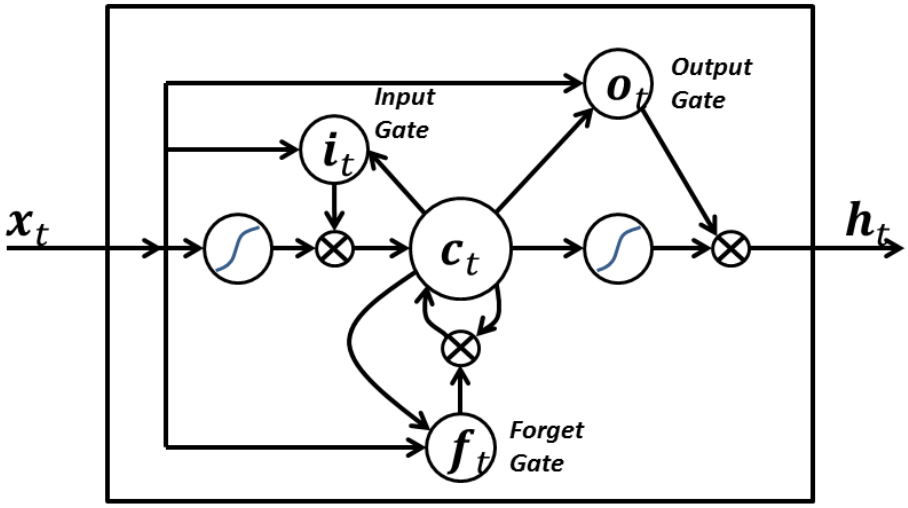


Figura 12 – Unidade de memória LSTM [Towards Data Science, 2020].

À medida que mais novas informações são inseridas na célula LSTM, sua arquitetura permite que ele avalie três pontos essenciais: i) quanto de memória é necessário preservar, que é uma informação atualizada no estado da célula; ii) quanto de memória deve ser atualizada em função das novas informações recebidas, que também é um dado a ser atualizado no estado da célula; iii) quanto da memória atual deve ser lida para o próximo passo interativo de aprendizado. O hiperparâmetro que especificamos ao definir uma rede LSTM é o número de unidades de memória, também conhecido como tamanho da célula.

Existem diferentes arquiteturas LSTM disponíveis para implementação de redes neurais. Este trabalho fez uso de uma célula LSTM original. Neste tipo de célula existem dois portões: um aprende a dimensionar a ativação de entrada e o outro aprende a dimensionar a ativação de saída. Assim, a célula pode aprender quando incorporar ou ignorar novas entradas e quando liberar o recurso que representa para outras células. A entrada para uma célula é alimentada em todos os portões usando pesos individuais [Machine Intellegence, 2020].

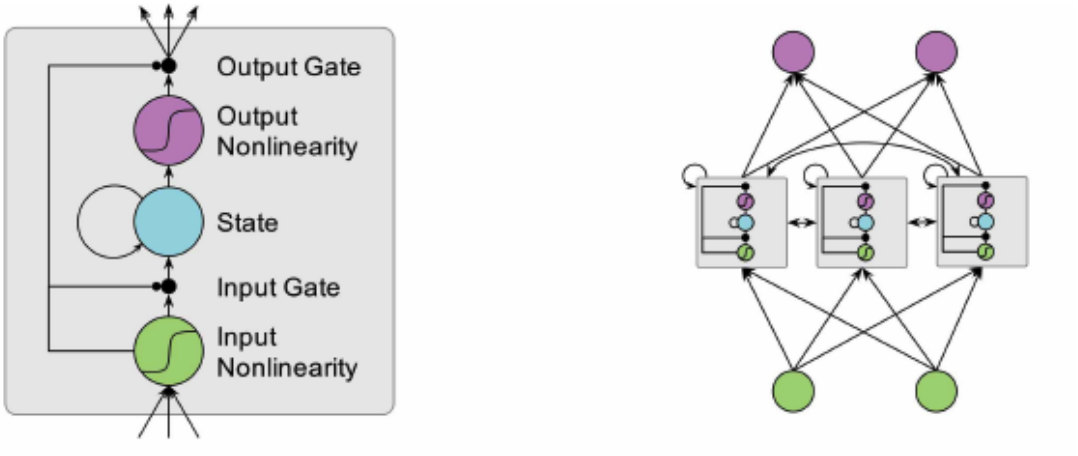


Figura 13 – Célula LSTM original

Uma extensão popular do LSTM tem como opção adicionar um portão de esquecimento que escala a conexão recorrente interna, permitindo que a rede aprenda a esquecer. Nesta extensão, a rede pode aprender a deixar o portão de esquecimento fechado, desde que seja importante lembrar o contexto da célula.

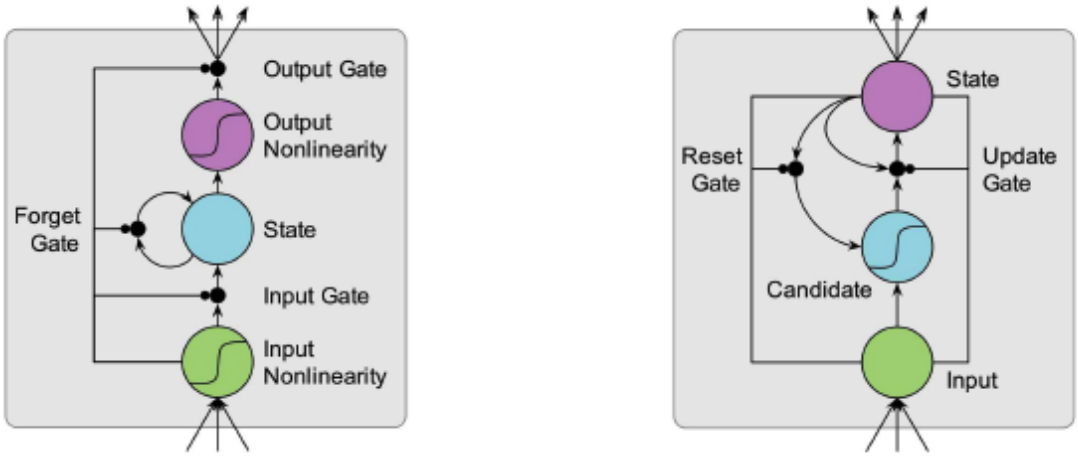


Figura 14 – Extensões de célula LSTM (Forget Gate e GRU)

Com base na ideia do LSTM, uma célula de memória alternativa chamada Unidade Recorrente Fechada (GRU) foi proposta em 2014 [Chung et al., 2014]. Ao contrário da LSTM, a GRU possui uma arquitetura mais simples e requer menos computação, enquanto produz resultados muito semelhantes. A GRU não possui porta de saída e combina as portas de entrada e esquecimento em uma única porta de atualização [Machine Intellegence, 2020].

1.2.7.4 Treinamento e testes

Os algoritmos tradicionais de aprendizado de máquina são categorizados em ”não supervisionados”, que exigem apenas dados de entrada, e “supervisionados”, que necessitam de dados de entrada e saída. As redes neurais profundas são uma forma de aprendizado supervisionado, e compreendem recentes avanços nos conhecimentos sobre aprendizado de máquina [Retson et al., 2019]. A figura 15 ilustra os algoritmos de aprendizado supervisionado e não-supervisionado.

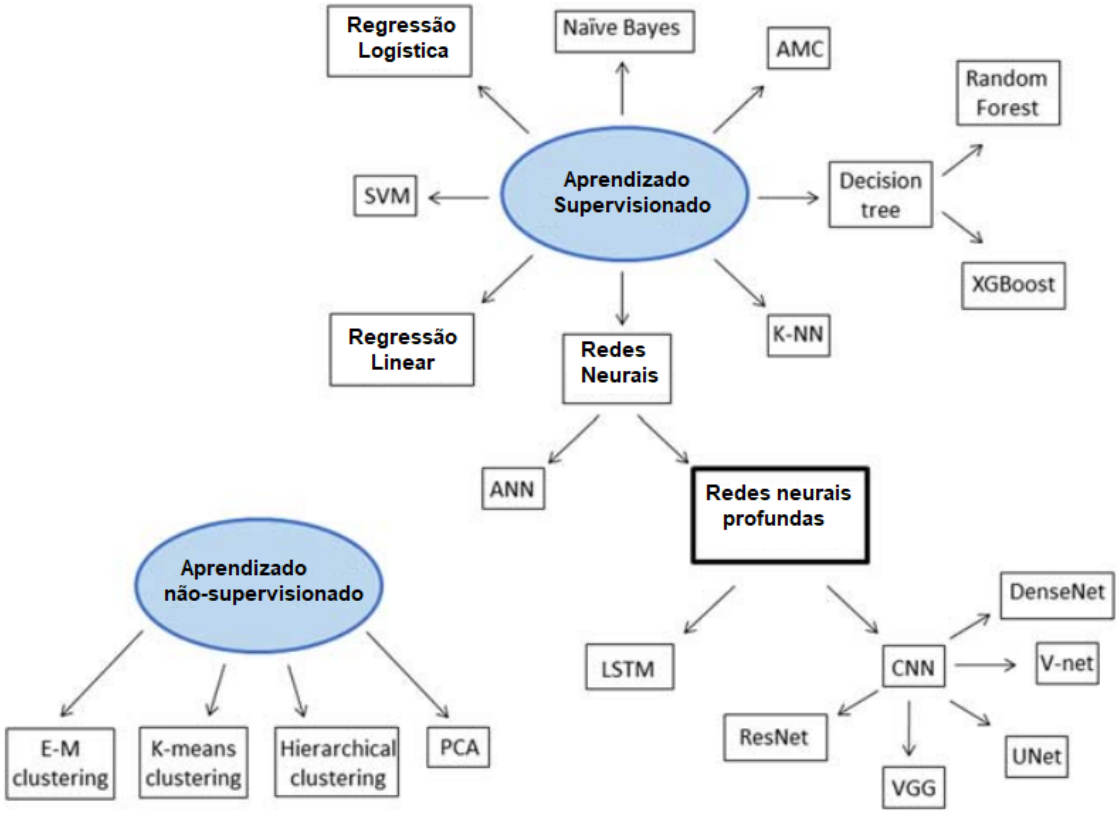


Figura 15 – Algoritmos de aprendizado de máquina, adaptado de [Retson et al., 2019]

O aprendizado supervisionado refere-se a técnicas nas quais um modelo é treinado em uma variedade de entradas, ou características, associadas a um resultado conhecido. Na medicina, isso pode representar o treinamento de um modelo para relacionar as características de uma pessoa (por exemplo, altura, peso, status de fumante) a um determinado resultado (início do diabetes em cinco anos, por exemplo). Depois que o algoritmo for treinado com sucesso, ele será capaz de fazer previsões de resultados quando aplicado a novos dados. As previsões feitas por modelos treinados usando aprendizado supervisionado podem ser discretas (por exemplo, positivas ou negativas, benignas ou malignas) ou contínuas (por exemplo, uma pontuação de 0 a 100) [Sidey-Gibbons and Sidey-Gibbons, 2019].

Para determinar o provável ponto de parada de treinamento, ou número de épocas, o qual apresente a melhor capacidade de generalização da rede, é utilizado o método de parada antecipada, ilustrado na figura 16, com base nas informações da técnica de validação cruzada. Esta abordagem é uma forma de validação que avalia um modelo de dados diferenciado do utilizado para treinamento, acompanhando os valores referentes a evolução do aprendizado, comparando ao longo da execução os valores de treinamento e validação. Deste modo, o treinamento pode ser interrompido ao se obter um valor mínimo do erro, antes do crescimento dos valores da função de validação [Haykin, 2009].

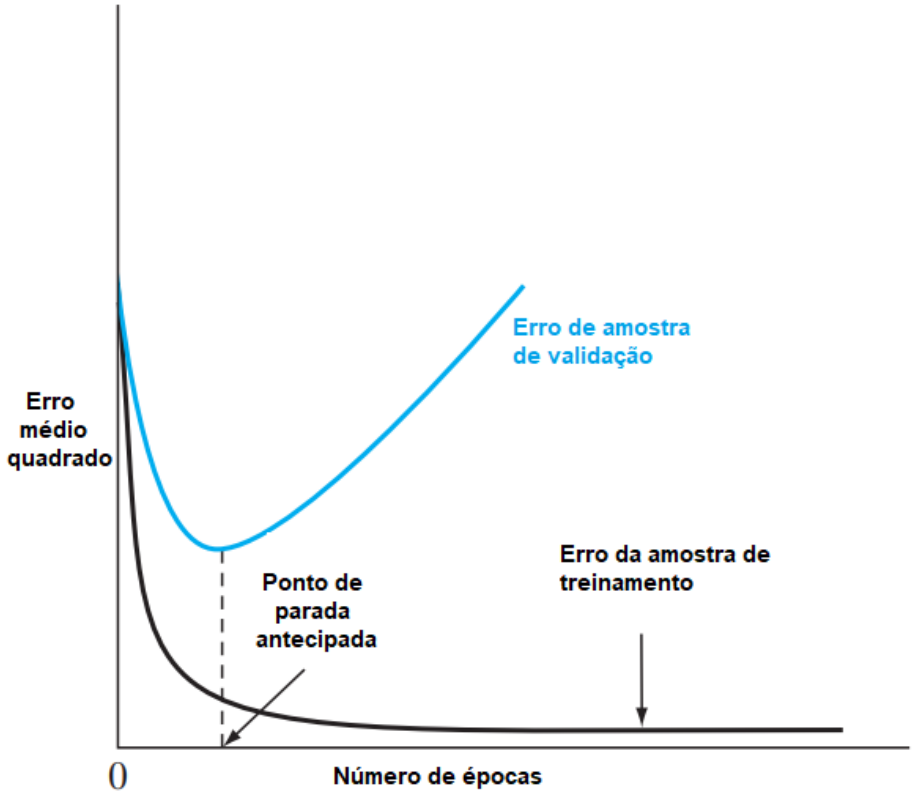


Figura 16 – Ilustração de regra de parada antecipada baseada na validação cruzada, adaptado de [Haykin, 2009]

Em termos de redes neurais artificiais, uma época refere-se a um ciclo no conjunto completo de dados de treinamento. Normalmente, o treinamento de uma rede neural leva mais do que algumas épocas. Em outras palavras, se alimentarmos uma rede neural com os dados de treinamento por mais de uma época em padrões diferentes, esperamos uma melhor generalização quando recebemos uma nova entrada “invisível” (dados de teste). Uma época geralmente é confundida com uma iteração. Iterações é o número de lotes ou etapas nos pacotes particionados dos dados de treinamento, necessários para concluir uma época. Heuristicamente, uma motivação é que (especialmente para conjuntos de treinamento grandes, mas finitos), ela oferece à rede a chance de ver os dados anteriores para reajustar os parâmetros do modelo, para que o modelo não seja influenciado pelos últimos pontos de dados durante o treinamento [DeepAI, 2020].

1.2.7.5 Métricas de avaliação

Um ponto necessário para compreensão dos resultados de execução de uma rede neural é o conhecimento das principais métricas de avaliação, utilizadas em modelos de classificação: acurácia, precisão, recall e f1, como ilustrado na figura 17.

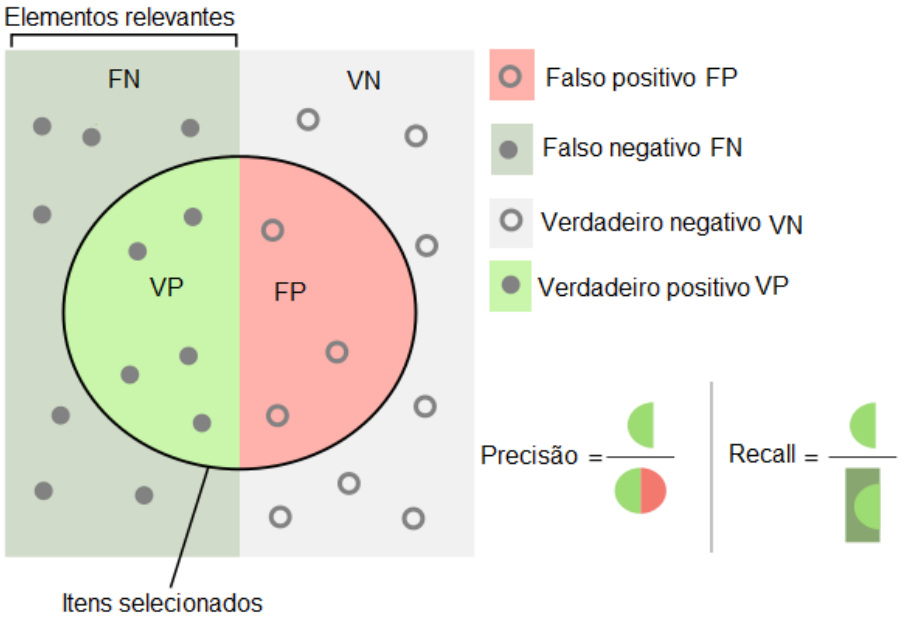


Figura 17 – Métricas de avaliação em modelos de classificação, adaptado de [Riggio, 2019]

- Acurácia: razão entre as observações previstas corretamente e o total de observações;

$$Acuracia = \frac{VP + VN}{VP + FP + FN + VN} \tag{1}$$

- Precisão: razão entrec o número de observações positivas previstas corretamente e o total de observações positivas previstas;

$$Precisao = \frac{VP}{VP + VN} \tag{2}$$

- Recall: proporção de observações positivas previstas corretamente para todas as observações na classe principal pesquisada;

$$Recall = \frac{VP}{VP + FN} \tag{3}$$

- F1: é a média harmônica entre Precisão e Recall; é amplamente usada para avaliar o sucesso de um classificador binário quando uma classe é rara [Lipton et al., 2014].

$$F1 = 2 * \frac{Precisao * Recall}{Precisao + Recall} \tag{4}$$

No capítulo 3 serão detalhados os procedimentos da metodologia, como a criação da base de dados, seu modelo de entidades e relacionamentos, o conjunto de ferramentas que gerou as informações do banco de dados, e os procedimentos para predição de proteínas essenciais, via rede neural.

2- Trabalhos relacionados

Este capítulo descreve os trabalhos relacionados utilizados no desenvolvimento desta dissertação. Foram pesquisados trabalhos abordando anotação funcional com base em ontologia, integração de dados, análise de similaridade semântica de proteínas, e predição de proteínas com uso de redes neurais.

2.1- Anotação funcional baseada em ontologia

A anotação de sequências genômicas da bactéria *E. coli* é considerada um dos trabalhos pioneiros na área, antes mesmo do sequenciamento do seu genoma estar disponível. Os trabalhos de classificação funcional desta bactéria ocorreram a partir de 1980, com base em suas funções metabólicas, sendo revistos em 1993 por Monica Riley, que criou um novo sistema de classificação [Hu et al., 2009].

Pesquisadores de uma das principais comunidades de estudo desta bactéria, a Ecocyc [Karp et al., 2014], adotaram a Gene Ontology no ano de 2005, e desde então contribuem ativamente com informações para o Consórcio Gene Ontology [Keseler et al., 2009].

Outro projeto sobre anotação de proteínas, que é uma citação importante para este trabalho, é o projeto de anotação de sequências da bactéria *Pseudomonas aeruginosa* PAO1, pois esta bactéria é referência para estudo das bactérias da espécie *Pseudomonas aeruginosa*. O PseudoCAP, *P. aeruginosa Community Annotation Project*, foi um projeto recebido com entusiasmo pelos pesquisadores, que apresentaram inicialmente um total de 1741 anotações, uma contribuição voluntária considerável para um genoma contendo 5570 genes. A publicação da sequência completa do genoma da *P. aeruginosa* PAO1 ocorreu no ano 2000, e o PseudoCAP foi o primeiro projeto de anotação de genoma totalmente baseado na Internet, apoiado por uma comunidade de pesquisadores para a análise de um genoma de um organismo de vida livre [Winsor, 2004].

Visando aprimorar qualitativamente as anotações desta bactéria, os membros de

sua comunidade de pesquisa se organizaram em 2014, e produziram 3533 sequências curadas e anotadas via GO. Um ponto importante a observar foi o uso da ferramenta InterProScan para execução no processo de anotações via GO. Desde então, o banco de dados desta bactéria se encontra em processo permanente de atualização de informações, com foco na adequação ao uso dos termos da GO [Winsor et al., 2015].

2.2- Integração de dados

Uma visão ampla e esclarecedora sobre integração de dados, em projetos de bioinformática, pode ser verificada em [Lapatas et al., 2015]. Este texto aborda temas variados de forma sucinta, que vão desde processos de compartilhamento de dados, anotações, bancos de dados públicos, até o detalhamento sobre os formatos de arquivos mais utilizados para troca de informações, entre diversos pontos, cujo conhecimento é essencial a quem se propõe a participar de projetos desta natureza.

Wanichthanarak et al. [2015] discutem estratégias para integração de dados genômicos, cujo interesse ao presente trabalho está na citação de importantes conjuntos de ferramentas, classificadas em categorias, voltadas a tarefas essenciais, como a análise de vias metabólicas, além de extenso ferramental em ambiente R sobre os assuntos, nas referências citadas.

2.3- Similaridade semântica

Em um trabalho pioneiro, Lord et al. [2002] descreveram uma investigação detalhada de medidas de similaridade semântica, examinando propriedades da GO. Chen et al. [2013] pesquisaram interações proteicas entre seres humanos e outros organismos. Este trabalho avaliou a capacidade de identificação de proteínas, a partir de 35 combinações de diferentes medidas de similaridade semântica entre termos da GO e produtos gênicos. Mais recentemente, Peng et al. [2018] apresentaram o método NETSIM2, o qual permite aos pesquisadores medir similaridades funcionais de genes com base na GO, conside-

rando a estrutura global da rede cofuncional com um método baseado na caminhada aleatória com reinício (RWR).

2.4- Predição de proteínas essenciais baseada em redes neurais

Redes neurais LSTM estão sendo empregadas em diferentes cenários de predição. O padrão LSTM de redes neurais é utilizado por Li et al. [2018], no qual afirma que as previsões baseadas em aprendizado de máquina das interações proteína-proteína podem fornecer informações valiosas sobre as funções das proteínas, a ocorrência de doenças e a identificação de tratamentos em larga escala. O trabalho de Hill et al. [2018] faz uso de redes neurais para a descoberta de regras biológicas complexas e para decifrar o potencial de codificação da proteína do RNA. Esta abordagem sugere o aprendizado de padrões complexos e de longo alcance em transcrições humanas completas, tornando-os ideais para executar uma ampla gama de tarefas difíceis de classificação e, o mais importante, para coletar novos conhecimentos biológicas, em função do crescente volume de dados de sequenciamento.

Outra característica interessante é a possibilidade de integração entre modelos diferenciados de redes neurais, como descrito no trabalho de Long et al. [2018], no qual integraram uma rede neural LSTM a outra de padrão CNN (*Convolutional Neural Network*), criando um modelo híbrido para predição de proteínas. Outra aplicação em bioinformática é verificada no trabalho de Metwally et al. [2019], como a pesquisa e predição de alergias alimentares a partir de perfis taxonômicos do microbioma intestinal. Alergias alimentares são difíceis de diagnosticar no início da vida, e podem levar a graves complicações de saúde ao indivíduo adulto, e o diagnóstico precoce é um aliado para tratamentos.

2.5- Considerações

Este trabalho se diferencia dos demais por apresentar uma abordagem computacional genérica, para análise das proteínas de um organismo, com apoio de ferramentas e

processos comparativos com proteomas de referência. A proposta do trabalho é fornecer informações funcionais sobre as proteínas do organismo em estudo, com a possibilidade de inferir quais destas proteínas têm perfil essencial.

A abordagem proposta combina resultados obtidos via ferramentas de análises genômicas, com os obtidos por técnicas de análise exploratória de dados, com base em aprendizado de máquina. Este processo analisa as sequências de proteínas do novo organismo, buscando inferir quais possuem perfil essencial, cujo conhecimento auxilia na identificação de novos fármacos.

O uso de técnicas de aprendizado de máquina, como a rede neural proposta neste trabalho, não é um recurso novo em análise de dados, mas que tem se aprimorado e sido utilizado em diversas áreas de conhecimento, inclusive em tarefas de bioinformática. Seu uso, aliado ao conhecimento disponível de ferramentas e bancos de dados biológicos, pode auxiliar a resolver problemas como a análise de sequências de proteínas, que demandam tempo, investimentos e recursos especializados de curadoria manual.

3- Metodologia

Este capítulo descreve a metodologia utilizada neste trabalho. A figura 18 ilustra os processos necessários para criação da base integrada de conhecimento sobre a bactéria *P. aeruginosa* CCBH4851. Cada processo da metodologia é descrito nas seções a seguir.

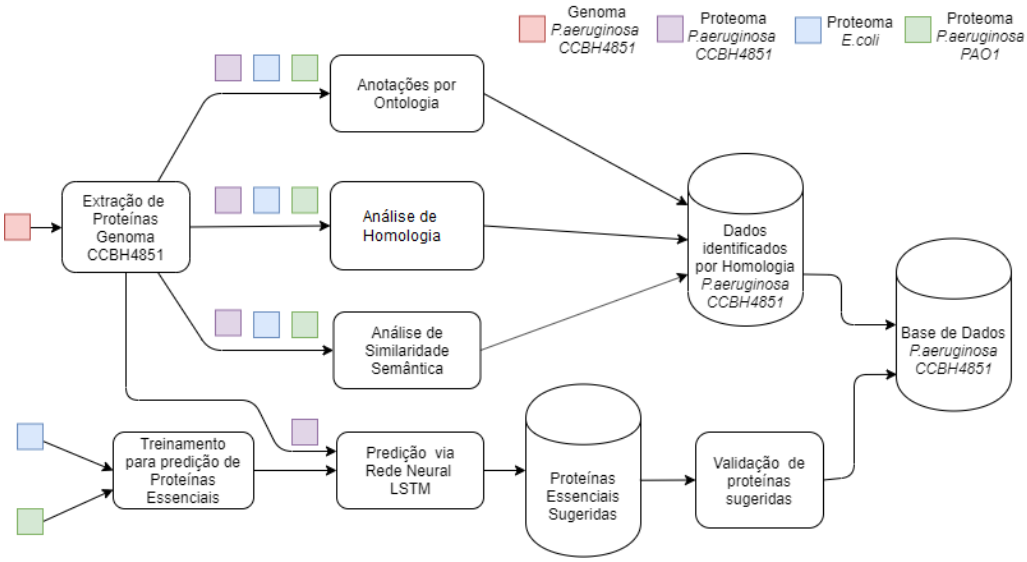


Figura 18 – Metodologia para construção da base de dados sobre a *P. aeruginosa* CCBH4851

3.1- Extração de proteínas

Este é o procedimento inicial da metodologia, que emprega o genoma da *P. aeruginosa* CCBH4851 para criação do arquivo de proteínas, conhecido como proteoma. Este processo consiste na leitura do arquivo do genoma e seleção e extração das sequências de proteínas contidas neste arquivo, no formato de arquivos Genbank [NCBI National Center for Biotechnology Information, 2018a].

A extração das sequências de proteínas, ou proteoma, da *P. aeruginosa* CCBH4851, foi efetuada por um *script* em linguagem Python, utilizando a biblioteca Biopython, dis-

ponível na página de arquivos do projeto[1]. Assim foi criado o arquivo de proteoma com base nas informações das regiões codificantes do genoma. A figura 19 exibe um extrato do genoma, informando um gene e sua região codificante (CDS, do inglês *coding sequence*).

```
gene            483..2027
                /locus_tag="PA4851_00005"
                /gene="dnaA"
CDS             483..2027
                /codon_start=1
                /gene="dnaA"
                /inference="COORDINATES: similar to AA
                sequence:RefSeq:NP_064721.1"
                /translation="MSVELWQQCVDLLRDELPSQQFNTWIRPLQVEAEGDELRVYAPN
                RFVLDWVNEKYLGRLLELLGERGEGQLPALSLLIGSKRSRTPRAAIVPSQTHVAPPPP
                VAPPPAPVQPVSAAPVVVPREELPPVTTAPSVSSDPYEPEEPSIDPLAAAMPAGAAPA
                VRTERNVQVEGALKHTSYLNRTFTFENFVEGKSNQLARAAAWQVADNLKHGYNPLFLY
                GGVGLGKTHLMHAVGNHLLKKNPNAKVVYLHSERFVADMVKALQLNAINEFKRFYRSV
                DALLIDDIQFFARKERSQEEFFHTFNALLEGGQQVILTSDRYPKEIEGLEERLKSRFG
                WGLTVAVEPPELETRVAILMKKAEQAKIELPHDAAFFIAQRIRSNVRELEGALKRVIA
                HSHFMGRPITIELIRESLKDLLALQDKLVSIDNIQRTVAEYYKIKISDLLSKRRSRSV
                ARPRQVAMALSKELTNHSLPEIGVAFGGRDHTTVLHACRKIAQLRESDADIREDYKNL
                LRTLTT"
                /transl_table=11
                /product="chromosome replication initiator DnaA"
                /locus_tag="PA4851_00005"
```

Figura 19 – Exemplo de região codificante - CDS

Cada CDS no arquivo do genoma, agrupa diversas informações, em conformidade com o formato de arquivo GenBank, definido pelo National Center for Biotechnology Information (NCBI) [NCBI National Center for Biotechnology Information, 2018a]. Seguindo as nomenclaturas adotadas neste formato, foram utilizadas as variáveis abaixo, indicando como exemplo os elementos da figura 19, para referência de valores:

- identificadores aplicados a cada gene no genoma (/locus_tag= ”PA4851_00005”) [NCBI National Center for Biotechnology Information, 2018c];
- nome do produto genético associado a sequência; (/product=”chromosome..,”)
- a tradução das sequências de aminoácidos das proteínas (/translation=”MSV...”)

Estas informações compõem os dados do arquivo de proteoma, em formato fasta, criado pelo *script* BioPython, submetido às ferramentas de análise genômica. O formato fasta é um tipo de arquivos padrão para troca de informações sobre sequências genômicas, e é comumente utilizado como parâmetro de entrada por ferramentas que analisam estas sequências.

[1] https://github.com/rjrmatias/cefet-rj/blob/master/arquivos/genbank2fasta.py

3.2- Anotação funcional baseada em ontologia

Para anotar funcionalmente as proteínas da *P. aeruginosa* CCBH4851, foram utilizadas as ferramentas Blast2GO [Gotz et al., 2008] e InterProScan [Jones et al., 2014]. Os processos das ferramentas utilizam bancos de dados como o *Gene Ontology*, Uniprot, Kegg e InterPro, tidos como referências em prover sequências genômicas curadas de proteínas de diversos microrganismos para estudos. A escolha das ferramentas Blast2GO e InterProScan se baseou no número de citações em projetos de anotação genômica.

A ferramenta InterProScan é de uso livre, enquanto a Blast2GO é comercial, dispondo uma versão básica com recursos limitados, de livre acesso, utilizada neste projeto. Esta versão, no entanto, já não é mais oferecida, em função de alterações no modelo de licenciamento e de uso da plataforma, e agora é parte integrante de um conjunto de ferramentas chamado OmicsBox [BioBam, 2019b].

A ferramenta InterProScan não oferece interface gráfica, operando por parâmetros informados via linha de comandos, exclusivamente no sistema operacional linux. A lista completa dos parâmetros da ferramenta pode ser acessada no endereço eletrônico [EMBL-EBI, 2019]. Foram informados os seguintes parâmetros para processamento:

- -i: arquivo de entrada, informado o arquivo do proteoma obtido;
- –goterms: aciona a busca das anotações Gene Ontology;
- –pa: aciona a busca de anotações de vias metabólicas;
- -iprlookup: valor de apoio ao processamento de vias metabólicas;
- -f JSON: especifica os formatos dos arquivos de resultados. O formato JavaScript Object Notation (JSON) foi utilizado neste estágio do projeto.

Ao término do processamento, a ferramenta cria o arquivo em formato JSON, contendo as anotações baseadas em ontologias, e as informações sobre vias metabólicas. A figura 20 ilustra uma ocorrência resultante deste processamento.

No exemplo em questão, foram identificados dois termos GO para a proteína, indicados no agrupamento *goXRefs*, GO:0008152 e GO:0008168. A ferramenta indica a proveniência dos dados, que é o banco de dados GO, e que somente ocorreram

```
"entry" : {
   "accession" : "IPR014777",
   "name" : "4pyrrole_Mease_sub1",
   "description" : "Tetrapyrrole methylase, subdomain 1",
   "type" : "HOMOLOGOUS_SUPERFAMILY",
   "goXRefs" : [ {
     "name" : "metabolic process",
     "databaseName" : "GO",
     "category" : "BIOLOGICAL_PROCESS",
     "id" : "GO:0008152"
   }, {
     "name" : "methyltransferase activity",
     "databaseName" : "GO",
     "category" : "MOLECULAR_FUNCTION",
     "id" : "GO:0008168"
   } ],
   "pathwayXRefs" : [ {
     "name" : "Synthesis of diphthamide-EEF2",
     "databaseName" : "Reactome",
     "id" : "R-HSA-5358493"
   } ]
 }
```


Figura 20 – Anotação Gene Ontology via InterProScan

anotações nas categorias processo biológico e função molecular, não ocorrendo anotação para componente celular.

A ferramenta Blast2GO era oferecida nas versões básica e comercial, e embora contasse com um número reduzido de funcionalidades, a versão básica procedeu com a anotação das proteínas, e forneceu mais termos que a ferramenta InterProScan. A anotação por ontologia na ferramenta Blast2GO é executada nas etapas Basic Local Alignment Search Tool (BLAST), mapeamento (Blast2GO Mapping), e anotações (Blast2GO Annotations), nesta ordem, como ilustrado na figura 21. Cada etapa depende dos resultados das etapas anteriores.

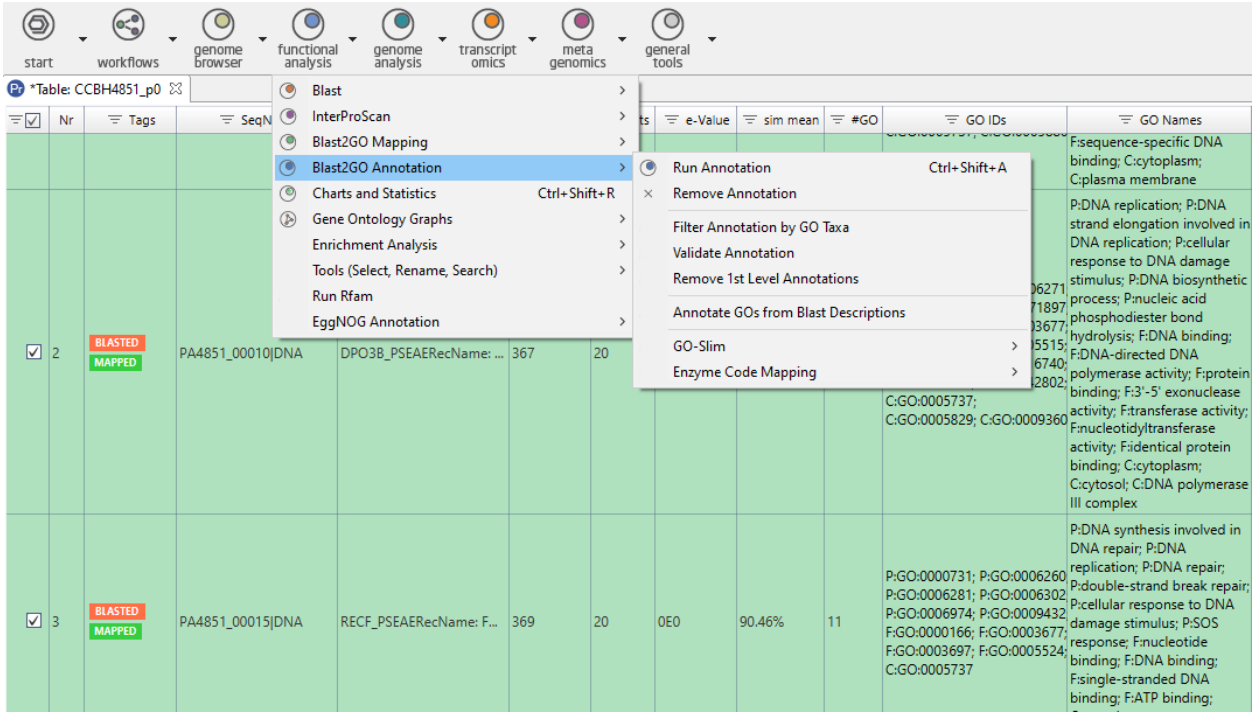


Figura 21 – Anotação de proteínas via Blast2GO

As opções do processo de anotação por ontologia, na ferramenta Blast2GO, efetuam os seguintes procedimentos:

- BLAST: localiza regiões de similaridade entre sequências. O programa compara sequências de nucleotídeos ou proteínas, com bancos de dados de sequências, e calcula a significância estatística das correspondências. O BLAST pode ser usado para inferir relações funcionais e evolutivas entre as sequências, além de ajudar a identificar membros de famílias de genes [Altschul et al., 1990].

- Mapeamento: é o processo de recuperar termos da GO, associados aos hits obtidos, após o processamento BLAST [BioBam, 2019a].

- Anotação: este processo seleciona os termos GO, a partir do conjunto de dados obtido na fase de Mapeamento, para atribuí-los às sequências de consulta. A anotação GO é realizada aplicando-se uma regra de anotação nos termos de ontologia encontrados. A regra procura encontrar as anotações mais específicas com um certo nível de confiabilidade [BioBam, 2019a].

3.3- Análise de homologia e ortologia

Este procedimento tem como objetivo avaliar e inferir quais proteínas da bactéria *P. aeruginosa* CCBH4851 podem ser ortólogas às proteínas das bactérias utilizadas como referência, *E.coli* e *P. aeruginosa* PAO1. Em função dos conhecimentos já adquiridos em estudos sobre estes organismos, é relevante identificar se há proteínas ortólogas, principalmente do ponto de vista funcional. Esta tarefa foi executada com apoio da feramenta Orthofinder [Emms and Kelly, 2015].

A ferramenta Orthofinder não possui interface gráfica, e funciona por parâmetros via linha de comandos, no sistema operacional linux. O único parâmetro utilizado foi a opção -f, que indica o nome do diretório no qual foram alocados os proteomas para análise. A figura 22 é um extrato dos grupos ortólogos identificados pela ferramenta, indicando a preservação entre as sequências de proteínas em destaque.

Orthogroup	CCBH4851	ecoli	pa01
OG0000030	PA4851_02480|acetyl-CoA, PA4851_27835|acetyl-CoA, PA4851_30875|acetyl-CoA	sp|P24182|ACCC_ECOLI	sp|P37798|ACCC_PSEAE, tr|Q9HTD0|Q9HTD0_PSEAE, tr|Q9I624|Q9I624_PSEAE
OG0000031	PA4851_03970|transcriptional, PA4851_10415|transcriptional, PA4851_13655|transcriptio	sp|P37641|YHJC_ECOLI	tr|Q9HWK7|Q9HWK7_PSEAE, tr|Q9HZX0|Q9HZX0_PSEAE, tr|Q9I0V0|Q9I0V0_PSEAE
OG0000032	PA4851_04080|branched-chain, PA4851_10050|lipase|COORDINATES_	sp|P37355|MENH_ECOLI,	tr|G3XCU7|G3XCU7_PSEAE, tr|Q9HWM9|Q9HWM9_PSEAE, tr|Q9HYA0|Q9HYA0_PSEAE

Figura 22 – Extrato de grupos ortólogos via Orthofinder

A análise por ortologia permite verificar comparativamente a preservação de trechos comuns nas sequências de proteínas, inferindo quais proteínas da *P. aeruginosa* CCBH4851 apresentam características funcionalmente semelhantes às das bactérias de referência, além de colaborar nas atividades de validação de dados.

3.4- Análise de similaridade semântica

A similaridade semântica entre as proteínas da *P. aeruginosa* CCBH4851 e as proteínas das bactérias *E.coli* e *P.aeruginosa* PAO1 foi analisada com o uso da ferramenta GOGO [Zhao and Wang, 2018]. Esta ferramenta compara termos *Gene Ontology* nos proteomas e infere um percentual de similaridade entre estes, como apresentado no exemplo da figura 23.

```
GO:1903097 GO:0010847 BPO 0.694
GO:0045252 GO:0045240 CCO 0.803
GO:0031216 GO:0004553 MFO 0.753
```

Figura 23 – Exemplo de processamento GOGO

Segundo seus desenvolvedores, a GOGO oferece vantagens por considerar métodos de análise híbridos e baseados em conteúdo de informações, como os métodos de Resnik [Resnik, 1999] e Wang [Wang et al., 2007]. Além disso, a ferramenta GOGO utiliza o indicador de Conteúdo de Informações (CI), mas não precisa recalculá-lo para um grande *corpus* de anotação de genes [Zhao and Wang, 2018]. Segundo du Plessis et al. [2011], o CI de um termo no grafo GO pode ser representado matematicamente pela expressão 5, onde *p*(c) é a probabilidade de ocorrência de um termo c no grafo.

$$ICF(c) = -log\ \boldsymbol{p}(c) \tag{5}$$

O termo raiz, que está implícito em todos os termos, tem probabilidade 1 e CI igual a 0. Por outro lado, termos raros têm um CI alto. A probabilidade do termo é geralmente estimada a partir de suas frequências, ou seja, o número de genes associados a c, dividido pelo número total de genes na ontologia [du Plessis et al., 2011].

3.5- Rede neural artificial

A rede neural proposta foi escrita em linguagem Python, utilizando as bibliotecas scikit-learn e keras, no ambiente Google Colab [Google, 2019]. Este ambiente é uma plataforma de desenvolvimento de softwares, de uso livre, que oferece acesso tanto a um amplo conjunto de bibliotecas pré-configuradas, quanto a um ambiente de execução e teste para estas rotinas, no qual o desenvolvedor pode inserir seus códigos e obter os resultados de execução. Estas facilidades são oferecidas por meio de notebooks Jupyter.

O propósito da rede neural foi inferir quais das proteínas da *P. aeruginosa* CCBH4851 poderiam ter perfil essencial. As proteínas essenciais e não-essenciais usadas para treinamento e teste foram obtidas a partir do banco de dados público OGEE. Foram selecionadas somente proteínas referentes às bactérias *E. coli* e *P. aeruginosa* PAO1.

3.5.1 Perfil das classes para treinamento e testes

A figura 24 descreve o perfil quantitativo das proteínas utilizadas dos organismos de referência.

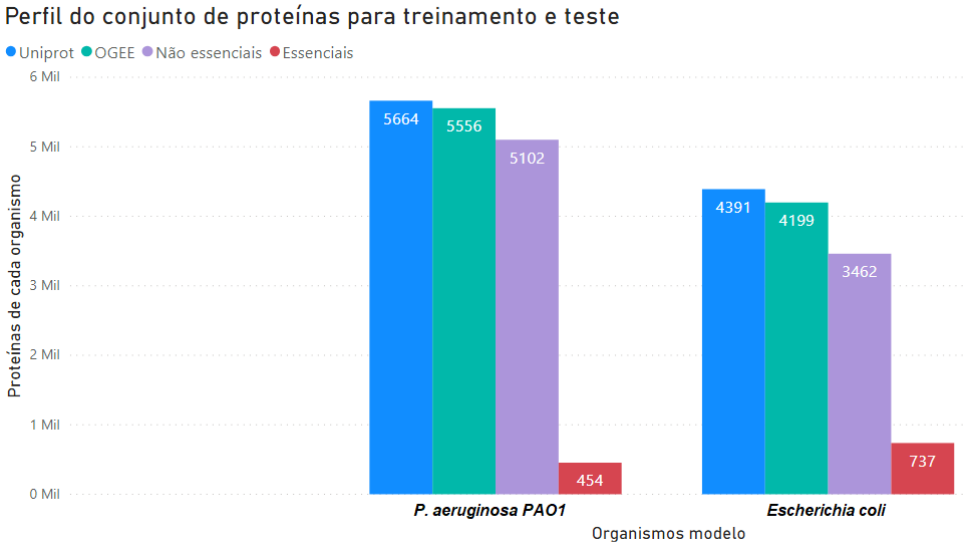


Figura 24 – Perfil quantitativo das proteínas utilizadas

A primeira coluna de cada organismo modelo, na cor azul, indica o total de proteínas proveniente do banco de dados genômico Uniprot. As colunas na sequência se referem ao total de proteínas fornecido pelo banco de dados OGEE, na cor verde, e em seguida as proteínas não-essenciais e essenciais, respectivamente, de cada organismo modelo referenciado. As proteínas essenciais correspondem a 8% do proteoma da cepa *P. aeruginosa* PAO1, e a 16% das proteínas da bactéria *E. coli*. O conjunto de proteínas totalizou 10354 proteínas.

Os índices de desbalanceamento entre as classes essencial e não-essencial podem ser observados na figura 24. Este índice entre as proteínas consideradas essenciais e não essenciais, das bactérias *P. aeruginosa* PAO1 e *E. coli*, é da ordem de 6,72% e 6,99%, respectivamente. Dados balanceados entre duas classes significaria valores próximos a 50% para cada uma. Para a maioria das técnicas de aprendizado de máquina, algum desequilíbrio não é um problema significativo. Porém, na ocorrência de casos, como 90% pontos para uma classe e 10% para a outra, critérios de otimização padrão ou medidas de desempenho podem não ser tão eficazes e precisariam ser modificados.

Deste modo, a estratégia inicial para analisar o conjunto de dados foi adotar a abordagem de execução por amostragem estratificada [Kim et al., 2013]. Neste tipo de amostragem, toda a população é dividida em estratos, ou subgrupos homogêneos, de acordo com o fator em estudo, no caso a característica essencial da proteína. Em seguida, são geradas amostras aleatórias dos diferentes estratos, que juntas representam a população analisada. As vantagens deste método são: i) permite que os pesquisadores obtenham um tamanho de efeito de cada estrato separadamente, como se fosse um estudo diferente. Portanto, as diferenças entre grupos se tornam aparentes e; ii) permite obter amostras de populações minoritárias/sub-representadas.

3.5.2 Arquitetura da rede

Este trabalho fez uso da arquitetura de rede neural LSTM. Redes neurais LSTM são orientadas a análise de dados em estruturas sequenciais, e dão suporte a padrões variáveis de entrada de dados, que são características das sequências de proteínas. O código-fonte 1 a seguir descreve a função de criação da rede neural.

Código 1 – Código de criação da rede neural

```
def create_lstm(embed_dim, lstm_out):
    model = Sequential()
    model.add(Embedding(max_features, embed_dim))
    model.add(SpatialDropout1D(0.4))
    model.add(LSTM(lstm_out, dropout=0.2, recurrent_dropout=0.2))
    model.add(Dense(2,activation='softmax'))
    model.compile(loss = 'binary_crossentropy', optimizer='adam',
    metrics=['acc'])
    return model
```

A arquitetura inicialmente proposta foi composta por quatro camadas, sendo a última de ativação por função Softmax. Esta função indica duas probabilidades, o quanto a proteína analisada se aproxima de ter características essenciais, e a probabilidade inversa, ou seja, sua proximidade de ser não-essencial.

A primeira instrução do código-fonte 1 define a configuração do tipo de modelo utilizado na rede neural como sequencial, que permite a inclusão de camadas em série na rede. A inclusão ocorre nas demais instruções com as diretivas *add*. A primeira camada acrescentada na rede, *Embedding*, é disponibilizada pela biblioteca Keras visando processamentos de texto. A camada acrescentada em seguida, *SpatialDropout1D*, é uma especialização da camada *Dropout*, que é uma técnica utilizada para prevenir a ocorrência de *overfitting* na rede. A camada final executa o processamento referente a função de ativação, do tipo *softmax*.

A função *softmax* calcula a evidência de um determinado registro pertencer a uma das classes avaliadas, convertendo este cálculo em probabilidades [Torres. Ai, 2018]. O código fonte desta rede e parâmetros estão disponíveis na plataforma colaborativa Colab do Google[2].

A biblioteca Keras aceita os seguintes argumentos para definição de uma rede LSTM:

- *units*: inteiro positivo, dimensionalidade do espaço de saída.

[2]https://bit.ly/2orwwBA

- *activation*: função de ativação a ser usada, cujo padrão é a tangente hiperbólica (tanh).
- *kernel_initializer*: inicializador para a matriz de pesos do *kernel*, usada para a transformação linear das entradas.
- *bias_initializer*: inicializador para o vetor de viés.
- *unit_forget_bias*: booleano. Quando inicializado com valor *True*, adiciona 1 ao viés do portão de esquecimento na inicialização. Configurá-lo como *True* também forçará *bias_initializer* = ”zeros”.
- *dropout*: decimal entre 0 e 1. É a fração das unidades a serem descartadas para a transformação linear das entradas.
- *recurrent_dropout*: decimal entre 0 e 1. É a fração das unidades a serem descartadas para a transformação linear do estado recorrente.
- *return_sequences*: booleano. Se deve retornar a última saída na sequência de saída ou a sequência completa.
- *return_state*: booleano. Se deve retornar o último estado, além da saída.
- *go_backwards*: booleano (padrão Falso). Se *True*, processe a sequência de entrada para trás e retorne a sequência invertida.
- *stateful*: booleano (padrão Falso). Se *True*, o último estado para cada amostra no índice i em um lote será usado como estado inicial para a amostra do índice i no lote a seguir.

3.5.3 Treinamento e teste da rede neural

A estratégia inicial adotada para treinamento e testes da rede neural consistiu na execução em 50 épocas, via plataforma Colab do Google [Google, 2019], indicadas via avaliação das curvas de aprendizado da rede. O número de épocas é um hiperparâmetro que define o número de vezes que o algoritmo de aprendizado vai atuar na totalidade do

conjunto de dados de treinamento. A execução de uma época indica que cada amostra no conjunto de dados de treinamento teve a oportunidade de atualizar os parâmetros internos do modelo [Brownlee, 2019].

Em função da identificação de desbalanceamento entre os quantitativos de cada classe, como citado previamente na subseção 3.5.1, verificou-se a necessidade de se considerar avaliar técnicas como *undersampling* ou *oversampling*, visando alcançar bons índices de precisão nas tarefas de predição de proteínas essenciais. Estas duas técnicas consistem em abordagens opostas: a técnica de *undersampling* consiste em reduzir o número de amostras da classe com maior volume de dados, no caso as proteínas não essenciais. Já a técnica de *oversampling* propõe ampliar o número de amostras das classes do menor conjunto, o que no presente trabalho seria o equivalente a replicar um quantitativo das proteínas essenciais, ampliando o conjunto destas proteínas.

3.6- Projeto da base de dados

Durante a fase de levantamento de requisitos para criação de um banco de dados, é necessário elaborar uma documentação descritiva, sobre as regras de negócio que o banco de dados visa atender. Este texto sucinto é conhecido como mini-mundo, e delimita o escopo de criação das entidades que vão ser descritas no modelo de dados. O texto a seguir descreve o mini-mundo para a base de dados da *P. aeruginosa* CCBH4851.

“Um conjunto de proteínas, ou proteoma, de um organismo em estudo, pode ser anotado por ontologia, por ferramentas genômicas. Uma ontologia pode descrever as características de uma proteína, que podem ter índices de similaridade semântica e de semelhança com outras proteínas, de organismos referência no estudo da espécie analisada. As proteínas do grupo de referência podem ser ortólogas às proteínas do organismo em estudo, e ambas pode estar associadas a um perfil essencial”.

A tabela 1 apresenta as descrições de cada entidade do modelo de dados elaborado. São informados o nome, a descrição e proveniência de dados.

Tabela 1 – Descrição das entidades do modelo de dados

Entidade	**Descrição**	**Proveniência**
ccbh	Proteoma da bactéria P. aeruginosa CCBH4851	Fiocruz
pao1	Proteoma da bactéria P. aeruginosa PAO1	Uniprot
ecoli	Proteoma da bactéria Escherichia coli	Uniprot
ccbh_blast2go	Anotações Blast2GO	Blast2GO
ccbh_interproscan	Anotações InterProScan	InterProScan
ccbh_bins_sim	Faixas de similaridade semântica	GOGO
ccbh_go_pao1	Termos Gene Ontology anotados entre as bactérias P. a. CCBH4851 e P. a. PAO1	Blast2GO e InterProScan
ccbh_go_ecoli	Termos Gene Ontology anotados entre as bactérias P. a. CCBH4851 e E. coli	Blast2GO e InterProScan
ccbh_pred	Predição de proteínas essenciais da bactéria P. a. CCBH4851	Rede neural LSTM
ecoli_ogee	Proteínas essenciais e não essenciais da bactéria E. coli	OGEE
pao1_ogee	Proteínas essenciais e não essenciais da bactéria P. a. PAO1	OGEE
ecoli_essencial	Proteínas essenciais da bactéria E. coli	OGEE
pao1_essencial	Proteínas essenciais da bactéria P. a. PAO1	OGEE
ecoli_blast_	Análise blast entre as bactérias P. a. CCBH4851 e E. coli	Blastp
pao1_blast_	Análise blast entre as bactérias P. a. CCBH4851 e E. coli	Blastp
ecoli_besthits	Blast mais significativo entre as bactérias P. a. CCBH4851 e E. coli	Blastp
pao1_besthits	Blast mais significativo entre as bactérias P. a. CCBH4851 e P. a. PAO1	Blastp
ortogrupos	Códigos e descrições dos grupos ortólogos verificados	Orthofinder
orto_all	Grupos ortólogos das bactérias P. a. CCHB4851, E. coli e P. a. PAO1	Orthofinder

O modelo de dados elaborado no SGBD MySQL, é apresentado na figura 25. Esta figura ilustra o modelo de entidades e relacionamentos da base de dados da *P. aeruginosa* CCBH4851. Este tipo de modelo foi desenvolvido para facilitar o desenvolvimento do banco de dados, permitindo a especificação de um esquema para representar a estrutura lógica geral de um banco de dados [Silberschatz et al., 2001]. As entidades descritas no modelo foram convertidas em tabelas, que armazenam as informações extraídas dos diversos procedimentos da metodologia.

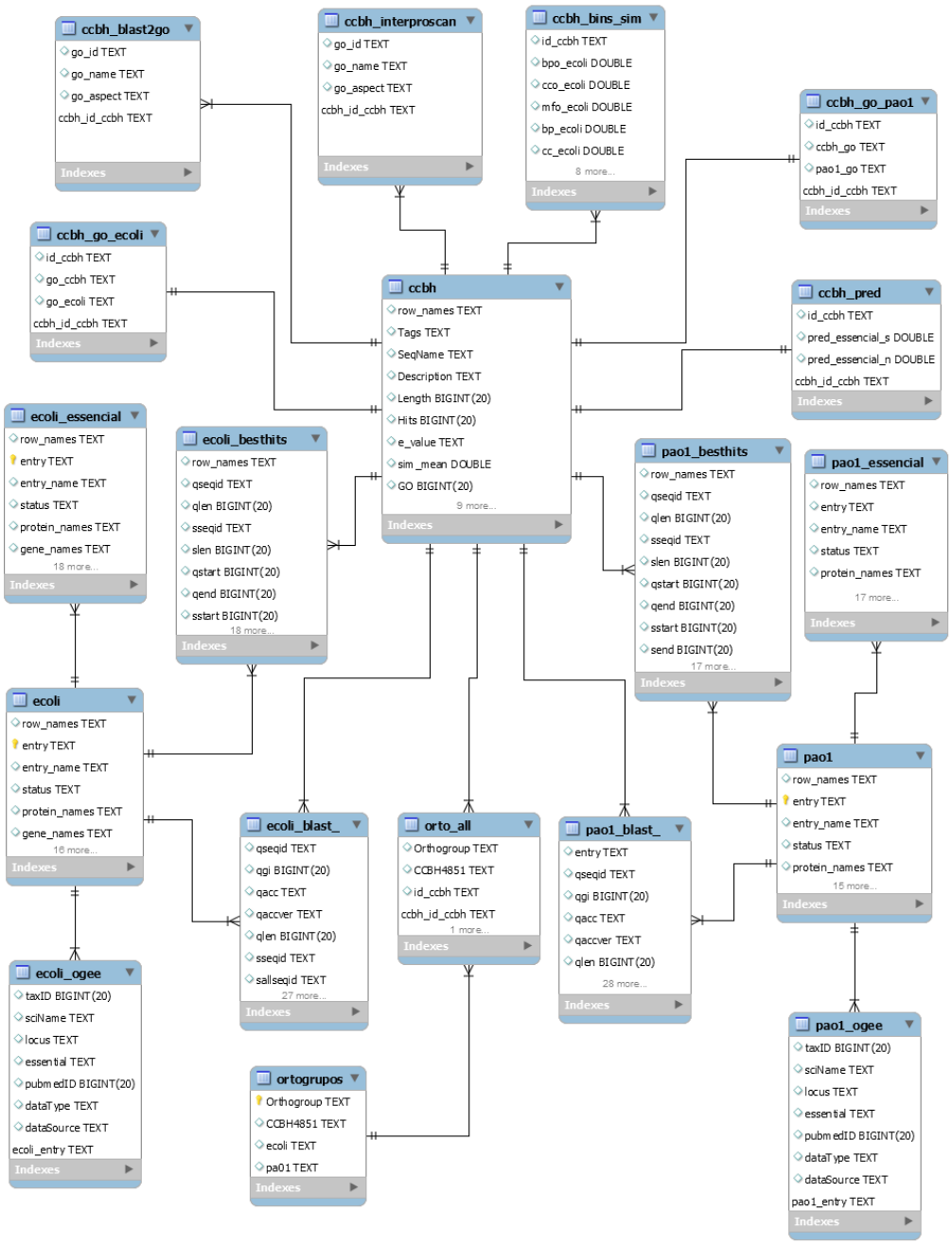


Figura 25 – Modelo de entidades e relacionamentos sobre a *P. aeruginosa* CCBH4851

3.7- Criação da base de dados

Um método comum de criação de um banco de dados consiste em executar *scripts*, escritos na linguagem Structured Query Language (SQL), com os comandos de definição das estruturas de dados que são criadas. Esta forma, no entanto, exige a escrita de todas as tabelas, com seus atributos e respectivos tipos de dados e dimensionamento, assim como as demais estruturas auxiliares, como índices e visões.

Existem outras possibilidades para criar um banco de dados. Um modo mais simples e eficiente, que o formato de execução de *scripts*, se baseia em transferir todas as definições das estruturas de dados, assim como todas as informações existentes nestas estruturas, diretamente ao SGBD. Este procedimento pode ser executado pela ferramenta RStudio [RStudio, Inc., 2018]. Esta ferramenta utiliza a linguagem R, que foi considerada a sétima linguagem mais utilizada pelos desenvolvedores no ano de 2018 [IEEE Spectrum, 2018], e que conta com uma ampla comunidade de desenvolvedores e usuários.

Esta abordagem foi utilizada em função de critérios como segurança, agilidade e praticidade. As rotinas de manipulação e transferência de dados, em plataformas como RStudio, são alvo de aprimoramento constante, ao mesmo tempo em que a linguagem R oferece facilitadores que permitem ao usuário, por meio de uma única linha de comando, transferir conjuntos de dados inteiros, com todos os seus atributos e valores, para formatos diferenciados de arquivos, ou gerenciadores de bancos de dados, como o MySQL.

Todos os resultados gerados nas etapas de processamento foram gravados em arquivos de formato texto, que são facilmente acessados pela ferramenta RStudio. Para gravação no SGBD MySQL, cada arquivo de resultados foi lido para estruturas do tipo *data frame* no RStudio, e posteriormente transferidos. Um *data frame* se assemelha a uma matriz, o qual pode ser referenciado por linhas e colunas, assim como ser utilizado em processos de transferência de dados. O código 2 exemplifica a carga e gravação de um arquivo em formato texto no SGBD MySQL.

Código 2 – Código de leitura e gravação de dados via RStudio no SGBD MySQL

```
con <- dbConnect(RMySQL::MySQL(),
                 dbname = "nome_do_banco_de_dados",
                 host = "servidor",
                 user = "usuario_do_banco_de_dados",
                 password = "senha_de_acesso")


ccbh <- read.csv(file="ccbh4851.txt",
                 header=TRUE,
                 sep="\t")

dbWriteTable(con,
             value = ccbh,
             name = "ccbh",
             append = TRUE,
             row.names = FALSE )
```

Todas as tabelas de dados do projeto foram gravadas seguindo este formato. Para executar este procedimento, o pacote *RMySQL* deve ser previamente instalado e carregado na ferramenta RStudio. O primeiro comando estabelece uma conexão com o MySQL, armazenando uma referência para esta conexão na variável *con*. Os parâmetros para a conexão, como o nome do banco de dados, o servidor em que o banco de dados está alocado, o usuário e senha de acesso devem ser informados.

Em seguida, o objeto *ccbh*, do tipo *data frame*, recebe integralmente o conjunto de dados referente ao arquivo *ccbh4851.txt*, com as informações do cabeçalho do arquivo, separados por tabulações. Por fim, o comando *dbWriteTable* efetua a criação da tabela e gravação dos dados contidos neste *data frame*, utilizando a conexão previamente estabelecida. Neste comando, são parametrizadas as diretivas de inclusão de dados, e de não numeração das linhas dos registros, na tabela recém-criada. Todas as informações do *data frame*, dados e metadados, que são as informações de cada atributo, são enviadas

ao SGBD. No capítulo seguinte serão apresentados os resultados alcançados, com os quantitativos refentes a cada etapa de processamento.

4- Resultados

As seções seguintes descrevem os resultados provenientes dos processos descritos na metodologia. Todos os resultados encontrados já integram a base de dados relacional, que serve como base de conhecimento preliminar para a cepa *P. aeruginosa* CCBH4851.

4.1- Extração de proteínas

O proteoma obtido a partir da execução do *script* Biopython totalizou 6211 proteínas. A figura 26 apresenta um extrato do proteoma indicando o nome de cada proteína, sua descrição e tamanho. Ao término da extração de proteínas, não é necessário executar qualquer outro procedimento no arquivo contendo o proteoma. O mesmo já se encontra pronto para ser utilizado como entrada de dados pelas ferramentas de anotação por ontologia.

Nr	SeqName	Description	Length
6182	PA4851_31450|hypothetical	protein|COORDINATES: similar to AA sequence:RefSeq:NP_254238.1	169
6183	PA4851_31455|bifunctional	glucosamine-1-phosphate acetyltransferase/N-acetylglucosamine-1-phosphate uridyltransferase|COORDINATES: similar to AA ...	454
6184	PA4851_31460|ATP	synthase subunit epsilon|COORDINATES: similar to AA sequence:RefSeq:NP_254240.1	141
6185	PA4851_31465|ATP	synthase subunit beta|COORDINATES: similar to AA sequence:RefSeq:NP_254241.1	458
6186	PA4851_31470|ATP	synthase subunit gamma|COORDINATES: similar to AA sequence:RefSeq:NP_254242.1	286
6187	PA4851_31475|ATP	synthase subunit alpha|COORDINATES: similar to AA sequence:RefSeq:NP_254243.1	514
6188	PA4851_31480|ATP	synthase subunit delta|COORDINATES: similar to AA sequence:RefSeq:NP_254244.1	178
6189	PA4851_31485|ATP	synthase subunit B|COORDINATES: similar to AA sequence:RefSeq:NP_254245.1	156
6190	PA4851_31490|ATP	synthase subunit C|COORDINATES: similar to AA sequence:RefSeq:NP_254246.1	85
6191	PA4851_31495|ATP	synthase subunit A|COORDINATES: similar to AA sequence:RefSeq:NP_254247.1	289
6192	PA4851_31500|ATP	synthase subunit I|COORDINATES: similar to AA sequence:RefSeq:NP_254248.1	126
6193	PA4851_31505|chromosome	partitioning protein|COORDINATES: similar to AA sequence:RefSeq:NP_254249.1	290
6194	PA4851_31510|chromosome	partitioning protein Soj|COORDINATES: similar to AA sequence:RefSeq:NP_254250.1	262
6195	PA4851_31515|16S	rRNA methyltransferase GidB|COORDINATES: similar to AA sequence:RefSeq:NP_254251.1	214
6196	PA4851_31520|tRNA	uridine 5-carboxymethylaminomethyl modification protein GidA|COORDINATES: similar to AA sequence:RefSeq:NP_254252.1	630
6197	PA4851_31525|hypothetical	protein|COORDINATES: similar to AA sequence:RefSeq:NP_254253.1	127
6198	PA4851_31530|hypothetical	protein|COORDINATES: similar to AA sequence:RefSeq:WP_019727085.1	117
6199	PA4851_31535|hypothetical	protein|COORDINATES: similar to AA sequence:RefSeq:WP_019727086.1	388
6200	PA4851_31540|integrase|COOR...	similar to AA sequence:RefSeq:WP_019727087.1	287

Figura 26 – Extrato do proteoma da *P. aeruginosa* CCBH4851

4.2- Anotação funcional baseada em ontologia

A anotação funcional é uma tarefa fundamental para determinar a função das proteínas durante a análise do proteoma [da Costa et al., 2018]. Este procedimento desempenha um papel chave, pois é a base de informações que integra todos os demais processamentos realizados. Resumidamente, a anotação de uma proteína infere dados relevantes, como sua localização, os processos biológicos em que participa, e sua função molecular. A figura 27 descreve os resultados obtidos nos processamentos de anotação funcional.

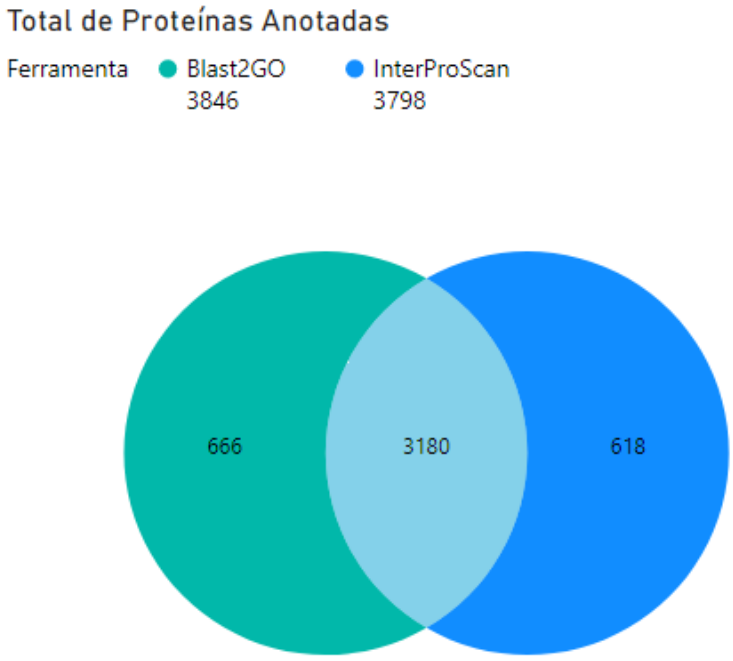
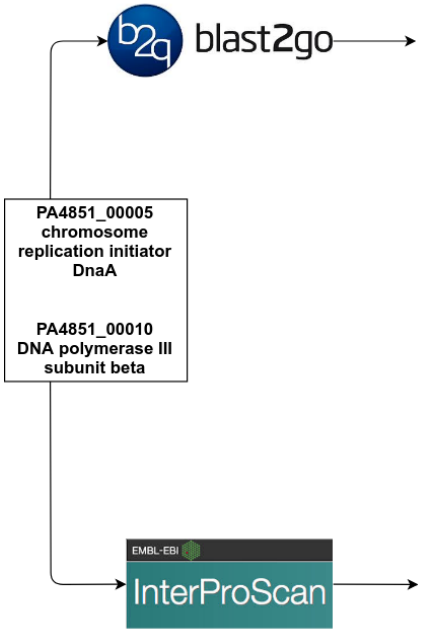


Figura 27 – Anotação de Proteínas via InterProScan e Blast2GO

Aproximadamente 60% das 6211 sequências do proteoma foram anotadas, sendo 3798 proteínas anotadas pela ferramenta InterProScan e 3846 proteínas anotadas pela ferramenta Blast2GO. Foram identificadas 3180 proteínas anotadas com pelo menos um termo semelhante.

A ferramenta Blast2GO apresentou resultados com diversidade de termos quantitativamente superior à ferramenta InterProScan. No conjunto de sequências com termos semelhantes, 1766 proteínas foram anotadas com pelo menos um termo idêntico, sendo

obtidos 2829 termos idênticos. Um total de 2413 sequências de proteínas foi anotado funcionalmente, e aproximadamente 45% deste total são descritas como proteínas hipotéticas. A figura 28 é um exemplo dos resultados referentes à anotação de duas proteínas, no qual verifica-se maior número de termos com a ferramenta Blast2GO, mas com ocorrência de resultados em comum para ambas.



Termo	id	Descricao	Categoria
GO:0003688	PA4851_00005	DNA replication origin binding	F
GO:0005524	PA4851_00005	ATP binding	F
GO:0005737	PA4851_00005	cytoplasm	C
GO:0005886	PA4851_00005	plasma membrane	C
GO:0006270	PA4851_00005	DNA replication initiation	P
GO:0006275	PA4851_00005	regulation of DNA replication	P
GO:0003677	PA4851_00010	DNA binding	F
GO:0003887	PA4851_00010	DNA-directed DNA polymerase activity	F
GO:0005829	PA4851_00010	cytosol	C
GO:0006271	PA4851_00010	DNA strand elongation involved in DNA replication	P
GO:0006974	PA4851_00010	cellular response to DNA damage stimulus	P
GO:0008408	PA4851_00010	3'-5' exonuclease activity	F
GO:0009360	PA4851_00010	DNA polymerase III complex	C
GO:0042802	PA4851_00010	identical protein binding	F
GO:0071897	PA4851_00010	DNA biosynthetic process	P
GO:0090305	PA4851_00010	nucleic acid phosphodiester bond hydrolysis	P

id	Termo	Descricao	Categoria
PA4851_00005	GO:0003677	DNA binding	F
PA4851_00005	GO:0003688	DNA replication origin binding	F
PA4851_00005	GO:0005524	ATP binding	F
PA4851_00005	GO:0006270	DNA replication initiation	P
PA4851_00005	GO:0006275	regulation of DNA replication	P
PA4851_00005	GO:0043565	sequence-specific DNA binding	F
PA4851_00010	GO:0003677	DNA binding	F
PA4851_00010	GO:0003887	DNA-directed DNA polymerase activity	F
PA4851_00010	GO:0006260	DNA replication	P
PA4851_00010	GO:0008408	3'-5' exonuclease activity	F
PA4851_00010	GO:0009360	DNA polymerase III complex	C

Figura 28 – Exemplo de anotação de proteínas via InterProScan e Blast2GO

Proteínas hipotéticas são aquelas em que se prevê serem expressas em um organismo, mas sem nenhuma evidência conhecida de sua existência [Ijaq et al., 2019]. As proteínas hipotéticas representam cerca de 28% de todo o proteoma da *P. aeruginosa* CCBH4851, um conjunto significativo de proteínas. No entanto, comparado ao percentual encontrado no proteoma de referência da *P. aeruginosa* PAO1, em que 40% de suas sequências são hipotéticas, é um valor plausível.

Foram encontradas 34 sequências de pseudogenes. Os pseudogenes são cópias

de genes que apresentam deficiências na sequência de codificação, como turnos de quadros e códons de parada prematura, mas que se assemelham a genes funcionais [Tutar, 2012]. As sequências de pseudogenes foram removidas do proteoma final, pois não se pode associar aos mesmos uma função, por não produzirem uma proteína funcional [Edwards et al., 2009].

Por fim, cada ferramenta utilizada produziu um total de termos correspondentes às três categorias *Gene Ontology*, descritos na figura 29.

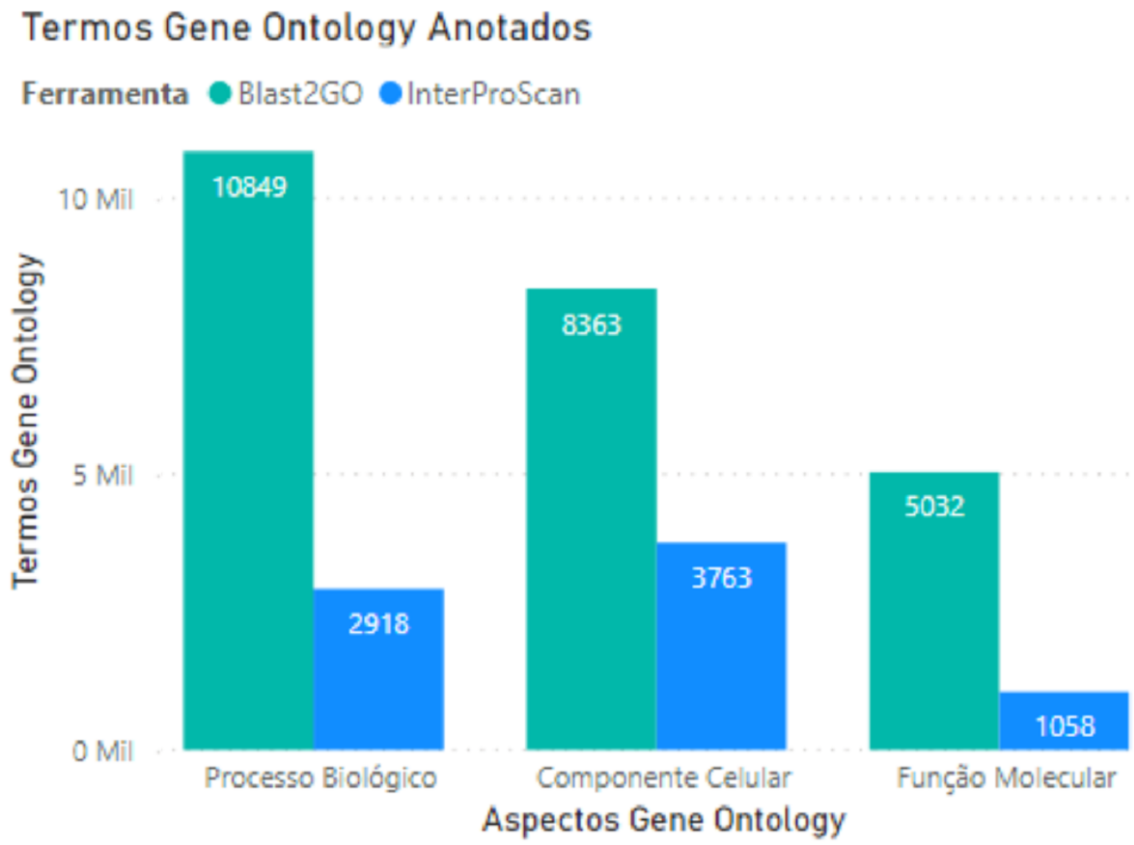


Figura 29 – Anotações Gene Ontology de acordo com as categorias

De modo a comparar os métodos de identificação de termos da GO, foi processada uma análise de concordância com base nos valores das categorias da GO obtidas por cada ferramenta, com uso do índice Kappa de Cohen [McHugh, 2012]. Esta análise se refere a capacidade de aferir resultados idênticos, aplicados ao mesmo sujeito/fenômeno, seja por instrumentos ou avaliadores diferentes, ou pelo mesmo instrumento em tempos diferentes, ou ainda por uma combinação entre estes fatores.

O índice Kappa de Cohen fornece uma medida de concordância entre dois métodos, como no cenário atual, em que há existência de duas ferramentas produ-

zindo resultados diferenciados, sobre o mesmo conjunto de dados. A concordância perfeita é evidente quando o Kappa de Cohen é igual a 1; um valor de Kappa de Cohen igual a zero sugere que o acordo não é melhor do que aquele que seria obtido apenas por acaso. Embora não exista escala formal, os seguintes níveis de concordância são frequentemente considerados apropriados [Watson and Petrie, 2010].

- Fraco se Kappa $\leq$ 0,00.
- Ligeiro se 0,00 $\leq$ Kappa $\leq$ 0,20.
- Justo se 0,21 $\leq$ Kappa $\leq$ 0,40
- Moderado se 0,41 $\leq$ Kappa $\leq$ 0,60
- Substancial se 0,61 $\leq$ Kappa $\leq$ 0,80
- Quase perfeito se Kappa $>$ 0,80

Deste modo, cada categoria identificada pela ferramenta InterProScan, foi comparada exclusivamente com o seu correspondente nos resultados da ferramenta Blast2GO. O índice resultante indicou ligeira correlação entre os resultados obtidos por cada ferramenta. Esta análise foi processada em linguagem R via RStudio, com o pacote estatístico "psych" [Revelle, 2018], e resultou nos valores descritos na tabela 2.

Tabela 2 – Análise de concordância das anotações *Gene Ontology*

Categoria	**Kappa**
Processo Biológico (PB)	0.159
Componente Celular (CC)	0.163
Função Biológica (FB)	0.132

4.3- Identificação de proteínas ortólogas

Este procedimento sugeriu um grupo significativo de proteínas, com características preservadas, observadas nos modelos de estudo avaliados. A figura 30 descreve o número total de proteínas de cada organismo modelo, bem como o número de proteínas ortólogas de cada modelo em relação ao organismo de estudo.

O processamento da ferramenta OrthoFinder sugeriu que 91% das proteínas da *P. aeruginosa* CCBH4851 são ortólogas às proteínas da *P. aeruginosa* PAO1. Do mesmo modo, 44% das proteínas da *P. aeruginosa* CCBH4851 são ortólogas as proteínas da bactéria *E. coli*. Foram identificadas 1934 proteínas ortólogas a ambos organismos de referência, totalizando aproximadamente 32% do total de suas proteínas, com características preservadas entre estes organismos.

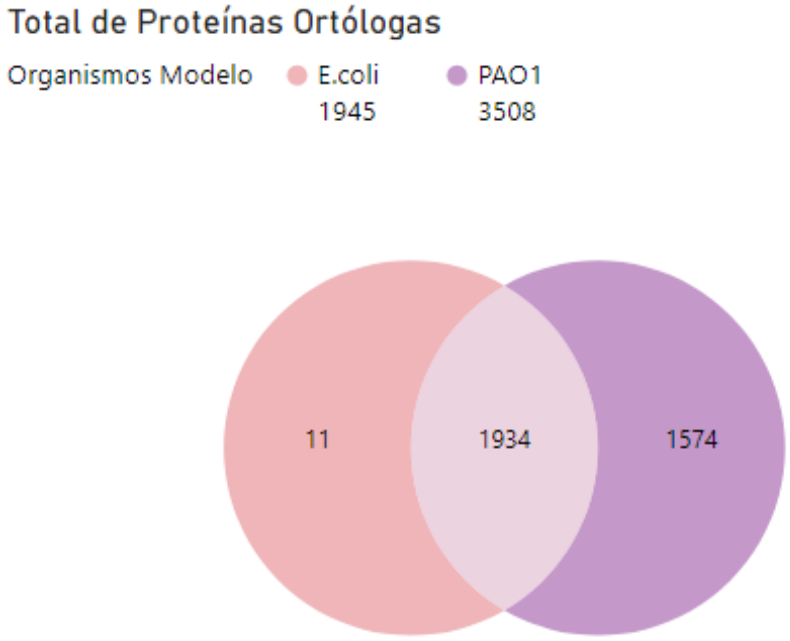


Figura 30 – Proteínas ortólogas

A partir do conjunto de proteínas resultante, foram obtidas as seguintes informações:

- 33% das proteínas são ortólogas às proteínas das bactérias *E.coli* e *P. aeruginosa* PAO1;
- 93% das proteínas ortólogas a todos os organismos foram anotadas via GO;
- 40% das proteínas ortólogas do organismo *P. aeruginosa* PAO1 não foram anotadas via GO;
- somente 11 proteínas ortólogas do organismo *E. coli* não foram anotadas via GO;

Esta análise relaciona as proteínas da bactéria *P. aeruginosa* CCBH4851 a ambos os organismos modelo. Considerando somente a bactéria *P. aeruginosa* PAO1, foi

observada preservação de características em praticamente todo o proteoma da bactéria *P. aeruginosa* CCBH4851, com percentual de 99,3% das proteínas inferidas como ortólogas.

O resultado inferiu um grupo de proteínas com características ortólogas, correspondendo a 44% das proteínas do modelo *E. coli* e 63% do modelo *P. aeruginosa* PAO1. Este resultado é um indicador que auxilia a inferir sobre as resultados de anotação funcional das proteínas. A figura 31 descreve um extrato dos ortogrupos inferidos. Neste exemplo, são listadas 11 proteínas da bactéria *P. aeruginosa* CCBH4851, para as quais ocorreu inferência de ortologia, para as três bactérias em conjunto. O ortogrupo OG0000037 indica que a proteína PA4851_08290 não encontrou correspondente com a bactéria *E. coli*, somente com a bactéria *P.aeruginosa* PAO1.

Orthogroup	CCBH4851	ecoli	pa01
OG0000030	PA4851_02480|acetyl-CoA, PA4851_27835|acetyl-CoA, PA4851_30875|acetyl-CoA	sp|P24182|ACCC_ECOLI	sp|P37798|ACCC_PSEAE, tr|Q9HTD0|Q9HTD0_PSEAE, tr|Q9I624|Q9I624_PSEAE
OG0000031	PA4851_03970|transcriptional, PA4851_10415|transcriptional, PA4851_13655|transcriptior	sp|P37641|YHJC_ECOLI	tr|Q9HWK7|Q9HWK7_PSEAE, tr|Q9HZX0|Q9HZX0_PSEAE, tr|Q9I0V0|Q9I0V0_PSEAE
OG0000032	PA4851_04080|branched-chain, PA4851_10050|lipase|COORDINATES_	sp|P37355|MENH_ECOLI,	tr|G3XCU7|G3XCU7_PSEAE, tr|Q9HWM9|Q9HWM9_PSEAE, tr|Q9HYA0|Q9HYA0_PSEAE
OG0000033	PA4851_05165|choline, PA4851_30120|choline, PA4851_30570|choline	sp|P0ABC9|BETT_ECOLI	tr|Q9HTI9|Q9HTI9_PSEAE, tr|Q9HTR3|Q9HTR3_PSEAE, tr|Q9HX83|Q9HX83_PSEAE
OG0000034	PA4851_05370|hypothetical, PA4851_08240|hypothetical, PA4851_21145|hypothetical	sp|P46482|AAEA_ECOLI	tr|Q9HXC1|Q9HXC1_PSEAE, tr|Q9HYU0|Q9HYU0_PSEAE, tr|Q9I4A7|Q9I4A7_PSEAE
OG0000035	PA4851_07040|nucleotide, PA4851_07135|GDP-mannose, PA4851_16840|UDP-glucose	sp|P76373|UDG_ECOLI	sp|O86422|UDG_PSEAE, sp|P11759|ALGD_PSEAE, tr|Q9HY58|Q9HY58_PSEAE
OG0000036	PA4851_07080|bifunctional, PA4851_15480|biofilm, PA4851_30955|phosphomannose	sp|P24174|MANC_ECOLI	sp|P07874|ALGA_PSEAE, tr|Q9HTB7|Q9HTB7_PSEAE, tr|Q9I1N7|Q9I1N7_PSEAE
OG0000037	PA4851_08290|hypothetical, PA4851_29045|hypothetical, PA4851_29995|hypothetical		tr|Q9HTT3|Q9HTT3_PSEAE, tr|Q9HU93|Q9HU93_PSEAE, tr|Q9HYC3|Q9HYC3_PSEAE, tr|
OG0000038	PA4851_08520|transcriptional, PA4851_10810|transcriptional, PA4851_21590|transcriptior	sp|P0ACM9|YIHL_ECOLI	tr|Q9HYZ1|Q9HYZ1_PSEAE, tr|Q9I041|Q9I041_PSEAE, tr|Q9I4J3|Q9I4J3_PSEAE
OG0000039	PA4851_10155|hypothetical, PA4851_11645|hypothetical, PA4851_18870|hypothetical	sp|P0AG38|RHTC_ECOLI	tr|Q9HZR8|Q9HZR8_PSEAE, tr|Q9I0D2|Q9I0D2_PSEAE, tr|Q9I3A2|Q9I3A2_PSEAE
OG0000040	PA4851_10170|histidine, PA4851_23130|arginine/ornithine, PA4851_29405|ABC	sp|P07109|HISP_ECOLI	sp|O30506|AOTP_PSEAE, sp|Q9HZS1|HISP_PSEAE, tr|Q9HU32|Q9HU32_PSEAE

Figura 31 – Extrato de proteínas ortólogas

4.4- Similaridade semântica

A aplicação de medidas de similaridade semântica é uma via para análise de dados genômicos, em função da organização hierárquica estruturada dos termos GO [Ayllón-Benítez et al., 2018]. Os indicadores de similaridade semântica são relevantes no contexto deste trabalho, principalmente em função das proteínas anotadas na categoria GO função molecular, para posterior análise das proteínas essenciais da bactéria *P. aeruginosa* CCBH4851. Neste procedimento, foram utilizadas as ferramentas Blastp e GOGO, referenciando os proteomas das bactérias *E. coli* e *P. aeruginosa* PAO1.

As figuras 32 e 33 descrevem os resultados de similaridade semântica obtidos com os referidos proteomas e o proteoma da *P. aeruginosa* CCBH4851. Nestes resultados, as proteínas foram inferidas com percentuais de similaridade semântica acima de 75%.

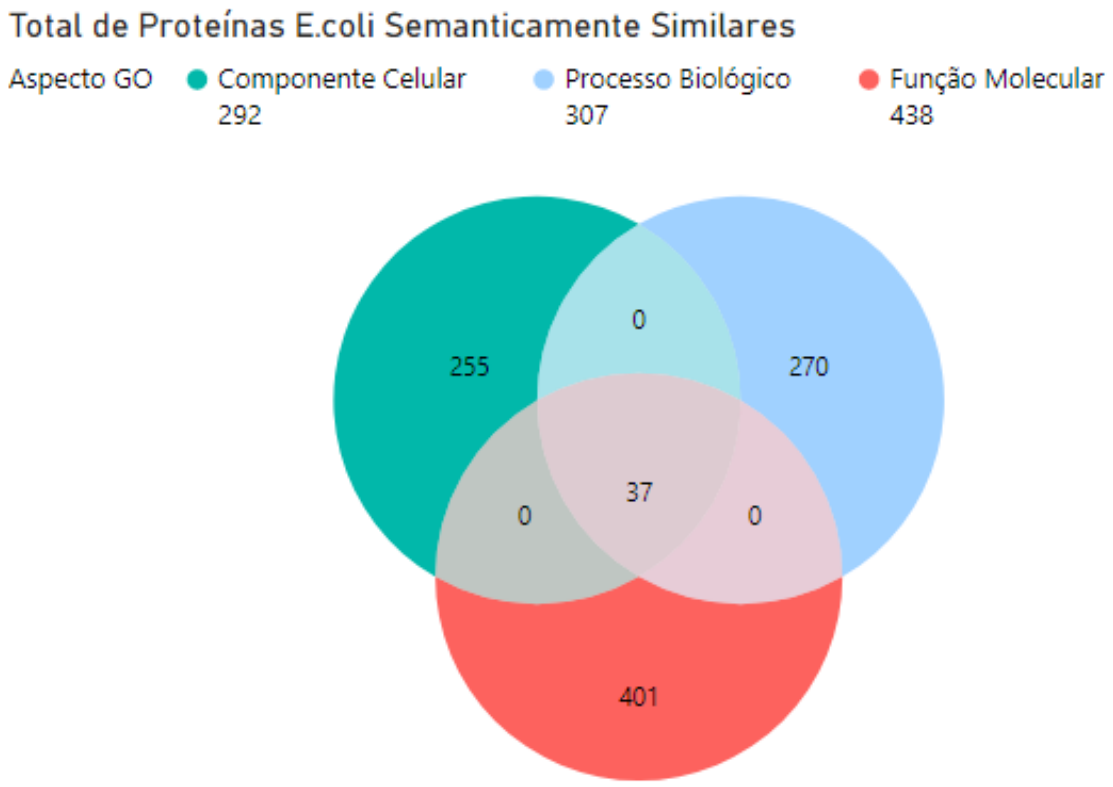


Figura 32 – Proteínas *E. coli* semanticamente similares

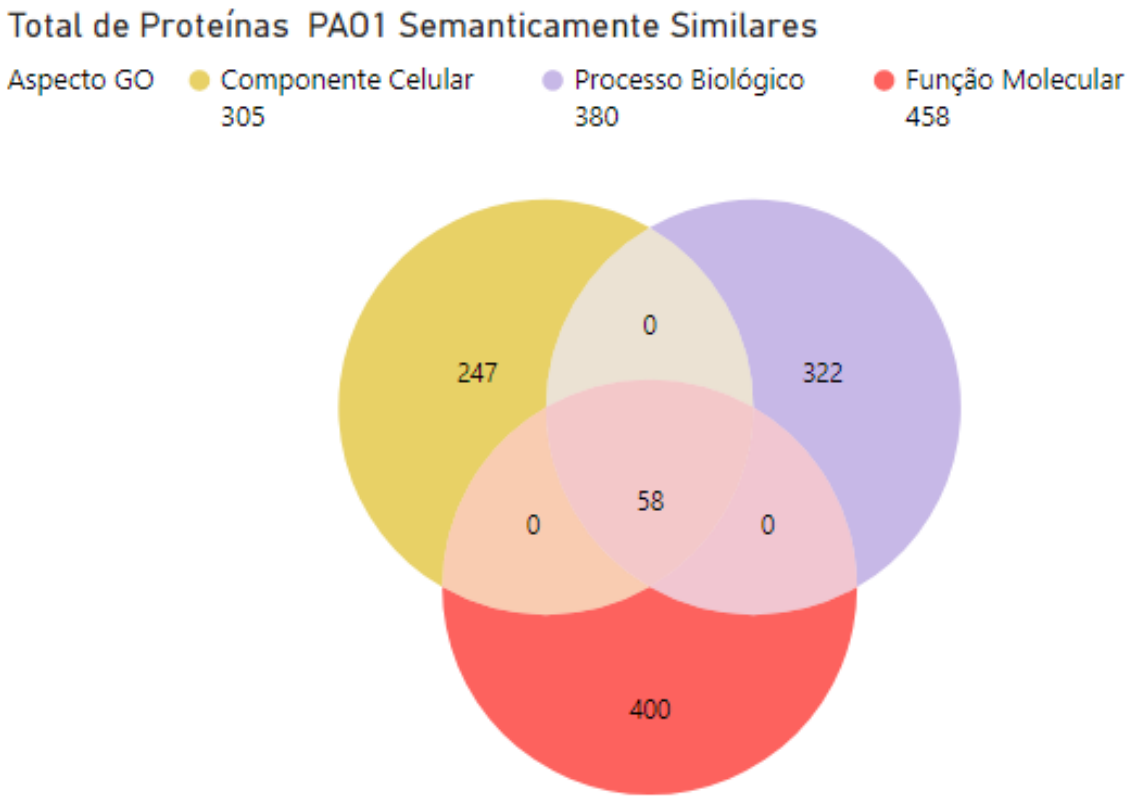


Figura 33 – Proteínas *P. aeruginosa* PAO1 semanticamente similares

A ferramenta GOGO inferiu o mesmo percentual, de 8% das proteínas, caracterizadas como altamente similares, para as bactérias *E. coli* e *P. aeruginosa* PAO1. Foram observadas 37 proteínas da bactéria *E. coli*, que englobam todas as categorias GO com o índice de 75%, e 58 proteínas da *P. aeruginosa* PAO1. O resultado cruzado, da análise de ambos os modelos, indicou as 14 proteínas apresentadas na figura 34, dentre as altamente similares em todos as categorias da GO, o que pode indicar um grupo relevante para estudos.

id	Descrição	Função Molecular	Processo Biológico	Componente Celular
PA4851_00070	SYGB_PSEA8Glycine--tRNA ligase beta subunit OS=Pseudomonas aeruginosa (strain L	91.6%	100.0%	100.0%
PA4851_01775	ILVD_PSEAEDihydroxy-acid dehydratase OS=Pseudomonas aeruginosa (strain ATCC 1	80.9%	100.0%	100.0%
PA4851_02255	YCIA_SALTYAcyl-CoA thioester hydrolase YciA OS=Salmonella typhimurium (strain LT2	88.6%	100.0%	100.0%
PA4851_02770	PGK_PSEAEPhosphoglycerate kinase OS=Pseudomonas aeruginosa (strain ATCC 156	80.1%	80.0%	100.0%
PA4851_05985	GUAA_PSEAEGMP synthase	79.6%	100.0%	100.0%
PA4851_08795	EDD_PSEAEPhosphogluconate dehydratase OS=Pseudomonas aeruginosa (strain ATC	80.6%	100.0%	100.0%
PA4851_09980	KTHY_PSEAEThymidylate kinase OS=Pseudomonas aeruginosa (strain ATCC 15692 / I	92.9%	76.7%	88.3%
PA4851_17820	NUDC_PSEAENADH pyrophosphatase OS=Pseudomonas aeruginosa (strain ATCC 15(	89.6%	100.0%	100.0%
PA4851_17985	SYC_PSEAECysteine--tRNA ligase OS=Pseudomonas aeruginosa (strain ATCC 15692 ,	89.4%	100.0%	88.3%
PA4851_20425	ACEK_PSEAEIsocitrate dehydrogenase kinase/phosphatase OS=Pseudomonas aerugir	75.8%	98.8%	100.0%
PA4851_23115	ARUC_PSEAESuccinylornithine transaminase/acetylornithine aminotransferase OS=Pse	76.3%	86.3%	100.0%
PA4851_28670	ILVE_PSEAEBranched-chain-amino-acid aminotransferase OS=Pseudomonas aerugino	100.0%	100.0%	100.0%
PA4851_30550	Y827_HAEINUncharacterized acyl-CoA thioester hydrolase HI_0827 OS=Haemophilus i	100.0%	100.0%	100.0%
PA4851_30820	PURK_PSEAEN5-carboxyaminoimidazole ribonucleotide synthase OS=Pseudomonas a	86.5%	100.0%	100.0%

Figura 34 – Proteínas inferidas como altamente similares e seus indicadores

Por fim, relacionado-se as proteínas com indicadores de similaridade semântica, contendo pelo menos um indicador na faixa percentual definida de 75%, com as proteínas ortólogas inferidas, foi verificada a ocorrência de 133 proteínas. Deste total, as 19 proteínas listadas na figura 35 têm similaridade semântica funcional acima de 75%.

id	Descrição	Função Molecular	Processo Biológico	Componente_Celular
PA4851_00070	SYGB_PSEA8Glycine--tRNA ligase beta subunit OS=Pseudomonas aeruginosa (strain LESB58	91.6%	100.0%	100.0%
PA4851_01340	DAVT_PSEAE5-aminovalerate aminotransferase DavT OS=Pseudomonas aeruginosa (strain AT	93.8%	84.8%	73.0%
PA4851_01775	ILVD_PSEAEDihydroxy-acid dehydratase OS=Pseudomonas aeruginosa (strain ATCC 15692 / [	80.9%	100.0%	100.0%
PA4851_02770	PGK_PSEAEPhosphoglycerate kinase OS=Pseudomonas aeruginosa (strain ATCC 15692 / DSI	80.1%	80.0%	100.0%
PA4851_03350	ARGC_PSEAEN-acetyl-gamma-glutamyl-phosphate reductase OS=Pseudomonas aeruginosa (s	80.5%	71.4%	100.0%
PA4851_05985	GUAA_PSEAEGMP synthase	79.6%	100.0%	100.0%
PA4851_06565	RRF_PSEAERibosome-recycling factor OS=Pseudomonas aeruginosa (strain ATCC 15692 / DS	100.0%	67.8%	88.3%
PA4851_08795	EDD_PSEAEPhosphogluconate dehydratase OS=Pseudomonas aeruginosa (strain ATCC 1569	80.6%	100.0%	100.0%
PA4851_08800	GLK_PSEAEGlucokinase OS=Pseudomonas aeruginosa (strain ATCC 15692 / DSM 22644 / CIF	80.3%	72.2%	100.0%
PA4851_09980	KTHY_PSEAEThymidylate kinase OS=Pseudomonas aeruginosa (strain ATCC 15692 / DSM 22	92.9%	76.7%	88.3%
PA4851_17820	NUDC_PSEAENADH pyrophosphatase OS=Pseudomonas aeruginosa (strain ATCC 15692 / DS	89.6%	100.0%	100.0%
PA4851_17985	SYC_PSEAECysteine--tRNA ligase OS=Pseudomonas aeruginosa (strain ATCC 15692 / DSM 2	89.4%	100.0%	88.3%
PA4851_20425	ACEK_PSEAEIsocitrate dehydrogenase kinase/phosphatase OS=Pseudomonas aeruginosa (str	75.8%	98.8%	100.0%
PA4851_21950	BRAF_PSEAEHigh-affinity branched-chain amino acid transport ATP-binding protein BraF OS=F	84.6%	80.9%	43.1%
PA4851_23115	ARUC_PSEAESuccinylornithine transaminase/acetylornithine aminotransferase OS=Pseudomor	76.3%	86.3%	100.0%
PA4851_28670	ILVE_PSEAEBranched-chain-amino-acid aminotransferase OS=Pseudomonas aeruginosa (strai	100.0%	100.0%	100.0%
PA4851_30285	ARGB_PSEAEAcetylglutamate kinase OS=Pseudomonas aeruginosa (strain ATCC 15692 / DSM	80.2%	59.5%	100.0%
PA4851_30550	Y827_HAEINUncharacterized acyl-CoA thioester hydrolase HI_0827 OS=Haemophilus influenza	100.0%	100.0%	100.0%
PA4851_30820	PURK_PSEAEN5-carboxyaminoimidazole ribonucleotide synthase OS=Pseudomonas aeruginos	86.5%	100.0%	100.0%

Figura 35 – Proteínas com similaridade semântica funcional acima de 75%

4.5- Treinamento e predição de proteínas essenciais

Antes de proceder com as tarefas de treinamento da rede neural, foi necessário executar uma etapa de limpeza de dados, para remover proteínas com tamanho desproporcional em relação às demais do conjunto. Uma análise preliminar com o conjunto de dados de treinamento, identificou proteínas cujo número de caracteres excedia substancialmente o tamanho médio das demais proteínas, e estas proteínas não tinham perfil essencial. Foi verificado ainda que sua existência no conjunto de dados depreciava o desempenho das tarefas de treinamento, teste, e predição. Deste modo, foram removidas 12 proteínas da bactéria *P. aeruginosa* PAO1 e uma proteína da bactéria *E. coli*.

Durante o treinamento da rede neural, foram obtidos os resultados de acurária, precisão, cobertura (*recall*) e F1, indicados na figura 36, onde as classes zero (0) e um (1) se referem às proteínas não-essenciais e essenciais, respectivamente.

	precision	recall	f1-score	support
0	0.89	0.52	0.66	1707
1	0.14	0.54	0.22	244
accuracy			0.53	1951
macro avg	0.51	0.53	0.44	1951
weighted avg	0.79	0.53	0.60	1951

Figura 36 – Medidas para avaliação de resultados de treinamento e teste

O processamento com a rede neural LSTM não proporcionou bons resultados. Considerando as medidas precisão e F1 para a classe 1, referente às proteínas essenciais, a execução retornou somente 14% e 22% respectivamente, como indicado na figura 36. Partindo destes resultados, foram testadas outras alternativas, como o uso das técnicas *oversampling*, com aumento do número de amostras da classe essencial, e posteriormente *undersampling*, com a redução gradual do conjunto de proteínas não-essenciais. Ainda assim não foram encontrados valores preditivos satisfatórios.

Para fins de análise exploratória, também foram avaliados outros classificadores, como *RandomForest* e *LogisticRegression*. Estas alternativas produziram resultados com valores muito próximos aos observados previamente com a rede LSTM. Neste sentido, foram feitas análises para verificar correlações não avaliadas. Os gráficos 37 e

38 descrevem a concentração de sequências essenciais e não-essenciais, em função do seu tamanho.

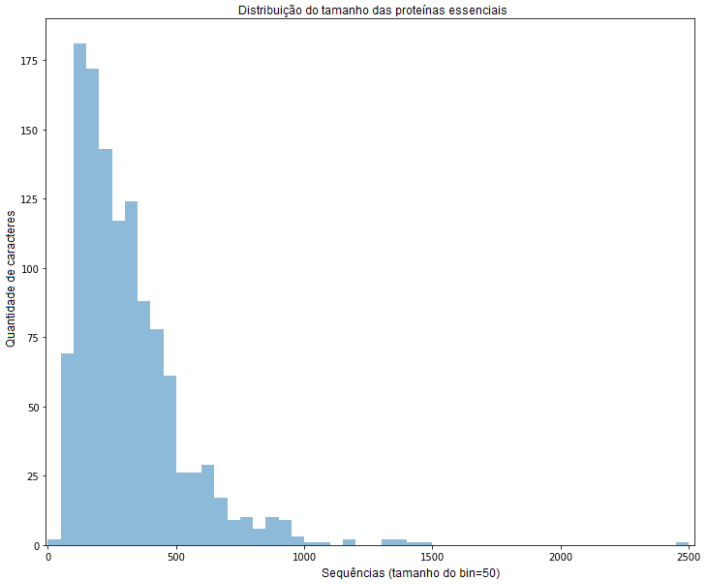


Figura 37 – Concentração de sequências essenciais por tamanho

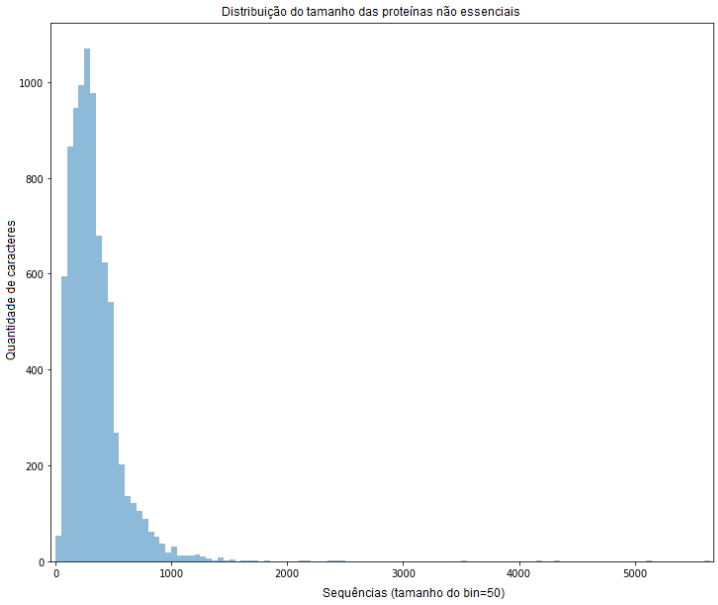


Figura 38 – Concentração de sequências não-essenciais por tamanho

Observou-se que a maioria das sequências essenciais contém até 180 caracteres, e as não-essenciais contém cerca de 600 caracteres. Buscou-se então analisar a correlação entre os conjuntos de ambas as classes, e o resultado indicado na figura 39 identifica um padrão muito próximo para os valores. O primeiro gráfico retrata o conjunto de proteínas essenciais, seguido pelo conjunto de proteínas não-essenciais.

Sobrepondo-se ambos, observa-se que não há diferenças significativas nos padrões de dados, o que dificulta a análise das sequências.

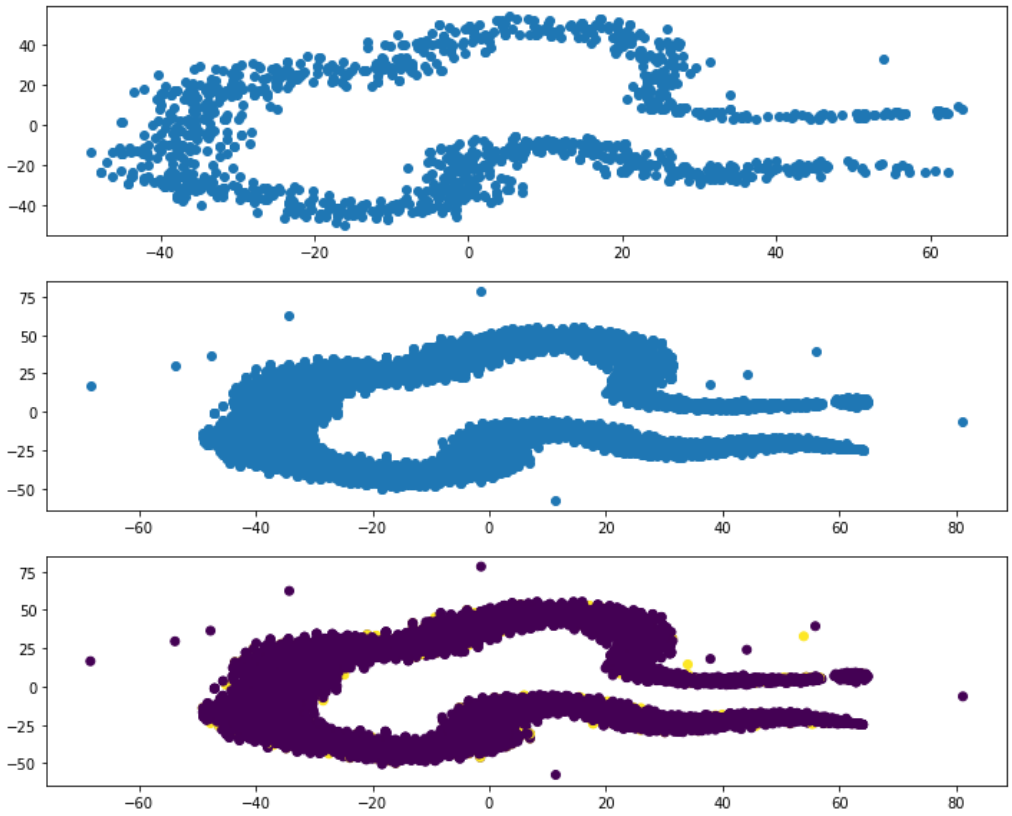


Figura 39 – Análise do padrão das proteínas essenciais e não-essenciais. Cada ponto no plot representa uma proteína, após redução de dimensionalidade pelo método t-SNE

Para visualização do padrão ilustrado na figura 39 foi utilizado o método de redução de dimensionalidade t-SNE [der Maaten, 2020]. Este método foi aplicado em função de cada proteína ser representada por um objeto multi-dimensional, antes representado por uma sequência de letras, passando a ser um objeto 2D após a redução. É importante observar que esta redução de dimensionalidade foi utilizada somente com a finalidade de avaliar o perfil das proteínas, pois ocorre perda de informações após a redução de um objeto de muitas dimensões para 2 dimensões.

Em função dos resultados insatisfatórios obtidos nas tarefas de treinamento e teste, observou-se que é necessário tanto aprimorar o modelo de predição, via rede neural LSTM, quanto ampliar o conjunto de características relacionadas às proteínas. Com o modelo atual não é possível inferir, com um grau elevado de confiabilidade, quais proteínas têm características essenciais ou não. Uma tarefa relevante para aprimoramento do modelo proposto consiste na identificação de um conjunto mais detalhado de características

sobre as proteínas, com intuito de auxiliar as tarefas de predição. O aprofundamento em características relacionadas às proteínas, obtidas a partir de estudos relacionados a redes metabólicas, motifs, interações proteína a proteína, reações enzimáticas, são alguns exemplos que podem auxiliar a aprimorar o modelo. Esta tarefa, no entanto, deve ser elaborada com a parceria de biólogos, capazes de sugerir quais são as características mais apropriadas que devem ser consideradas, para que o modelo alcance predições com maior grau de confiabilidade.

5- Considerações finais

O presente trabalho apresentou os direcionamentos utilizados para criar uma base de dados sobre a cepa *P. aeruginosa* CCBH4851. Esta base de dados já se encontra disponível publicamente via repositório Github, e contempla informações relevantes sobre as proteínas da bactéria, como as anotações pela ontologia Gene Ontology, os índices inferidos de similaridade semântica, e as sequências com características ortólogas. Os relacionamentos criados a partir dos resultados são facilitadores, que podem ser utilizados na descoberta de novas correlações, tanto entre os proteomas utilizados, quanto com novos organismos.

O processo de anotação por ontologia inferiu que aproximadamente 60% das proteínas da cepa *P. aeruginosa* CCBH4851 são similares a proteínas catalogadas no banco de dados UniProt. Em paralelo, análises comparativas elaboradas com os proteomas referência no estudo da espécie *Pseudomonas aeruginosa*, das bactérias *E. coli* e *Pseudomonas aeruginosa* PAO1, indicaram um grupo de proteínas com índices de similaridade semântica acima de 75%, com as proteínas da *P. aeruginosa* CCBH4851, que pode sinalizar um conjunto de sequências relevante para estudos.

Além dos valores de similaridade semântica e ortologia, a base de dados indica quais sequências têm maior similaridade de alinhamento, obtidas via processamento Blastp. Nesta análise, considerando percentuais acima de 75%, observou-se que a maior parte das proteínas da *P. aeruginosa* CCBH4851, cerca de 88% das sequências, se assemelham às proteínas da bactéria *Pseudomonas aeruginosa* PAO1, e aproximadamente 10% da *E. coli*. Estudos mais aprofundados sobre o perfil funcional destas proteínas podem conduzir a novos conhecimentos sobre a cepa *P. aeruginosa* CCBH4851.

Avaliando quantitativamente o total de anotações por ontologia obtido da cepa *Pseudomonas aeruginosa* CCBH4851, que compreende 63% das suas proteínas, foi observado que é inferior ao total de proteínas anotadas da cepa *Pseudomonas aeruginosa* PAO1, que se aproxima de 80%. O proteoma da cepa *Pseudomonas aeruginosa* PAO1 contém, no entanto, um número reduzido de proteínas, aproximadamente 10% menor, e com maior número de sequências hipotéticas, o que pode influir nesta análise.

Quanto às tarefas de predição de proteínas essenciais, por aprendizado de

máquina, foi observado que o procedimento precisa ser aprimorado, tanto no perfil do modelo empregado, quanto nas características utilizadas para as tarefas de predição, que não alcançaram índices de confiabilidade necessários para sugerir um grupo consistente de proteínas essenciais. Na busca por alternativas de solução, foram elaborados testes com diferentes técnicas e classificadores, que ainda indicaram resultados com baixa confiabilidade, reforçando a necessidade de rever o método empregado.

A metodologia deste trabalho baseou-se exclusivamente em processos computacionais. Embora sejam considerados menos confiáveis que os executados por curadoria manual, que demandam mais tempo e recursos especializados, as análises por processos computacionais fornecem resultados em prazos reduzidos. Os processos computacionais permitem que os resultados sejam constantemente aprimoradas por pesquisadores em nível global, além de permitir a análise de grandes volumes de dados, comuns em pesquisas científicas, e auxiliar nas análises preliminares de um organismo em estudo, como a *Pseudomonas aeruginosa* CCBH4851.

Um fator necessário para aprimorar a metodologia proposta, é a parceria com pesquisadores de campo, para comprovar via processos de curadoria manual, a acurácia das informações inferidas. A avaliação de profissionais que atuem diretamente nas análises dos genes da bactéria é um importante requisito para atestar a qualidade das informações disponibilizadas.

A metodologia proposta pode ser aplicada de modo genérico ao estudo de novos organismos, mesmo a genomas com um quantitativo elevado de sequências, e produzir resultados em curto espaço de tempo, que é um fator necessário, desejado e que auxilia na tomada de decisão dos pesquisadores. O processo proposto neste trabalho pode ser ampliado por trabalhos futuros, pois novos métodos analíticos de aprendizado de máquina, e técnicas envolvendo análise textual estão em alta demanda, e podem viabilizar novas soluções que são adaptáveis ao escopo de bioinformática, na análise de sequências de proteínas.

Para ampliar o acesso e favorecer a divulgação das informações da base de dados criada é necessário criar um canal de consulta a estes dados. Neste sentido, uma sugestão para trabalhos futuros seria a elaboração de uma interface gráfica web, e de um serviço web, ou *webservice*, disponíveis publicamente. A interface gráfica poderia oferecer ao usuário um conjunto de consultas sobre todo o contexto de informações explorado, e o *webservice* por sua vez, poderia oferecer as informações disponíveis

relacionadas as proteínas da *P. aeruginosa* CCBH4851, que poderiam ser acessadas tanto por sistemas externos quanto pela própria interface gráfica web.

A bactéria *P. aeruginosa* CCBH4851 é um organismo novo para estudos, que pertence a uma espécie de bactérias indicada pela OMS como prioritária para desenvolvimento de novos medicamentos. O conjunto de informações reunido na base de dados criada contribui para ampliar o conhecimento sobre esta cepa, e consequentemente sobre a espécie, e pode auxiliar nas tarefas em curso de modelagem de célula inteira, conduzidas pela Fiocruz.

Referências Bibliográficas

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., and Walter, P. (2014). *Molecular Biology of the Cell, Sixth Edition*. Taylor & Francis Group.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Ayllón-Benítez, A., Mougin, F., Allali, J., Thiébaut, R., and Thébault, P. (2018). A new method for evaluating the impacts of semantic similarity measures on the annotation of gene sets. *PLOS ONE*, 13(11):e0208037.

Berg, J. M., Tymoczko, J. L., Gatto, G. J., and Stryer, L. (2010). *Biochemistry*. W. H. Freeman and Company, 7 edition.

BioBam (2019a). *Blast2GO User Manual*. http://docs.blast2go.com/user-manual/. acessado em 09/11/2019.

BioBam (2019b). *OmicsBox: Bioinformatics Made Easy*. https://www.biobam.com/omicsbox/. acessado em 10/10/2019.

Birney, E. (2004). Biological database design and implementation. *Briefings in Bioinformatics*, 5(1):31–38.

Brennan, S. (2019). *Homology*. http://brennan564s17.weebly.com/homology.html. acessado em 05/11/2019.

Britz, D. (2016). *Recurrent Neural Networks Tutorial, Part 1 – Introduction to RNNs*. http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/. acessado em 27/10/2019.

Brown, T. A. (2002). *Genomes, 2nd edition*. Wiley-Liss, Washington, DC.

Brownlee, J. (2019). *Difference Between a Batch and an Epoch in a Neural Network*. https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/. acessado em 23/11/2019.

Campos, T. L., Korhonen, P. K., Gasser, R. B., and Young, N. D. (2019). An evaluation of machine learning approaches for the prediction of essential genes in eukaryotes using protein sequence-derived features. *Computational and Structural Biotechnology Journal*, 17:785–796.

Chatterjee, C. C. (2019). *Implementation of RNN, LSTM, and GRU*. https://towardsdatascience.com/implementation-of-rnn-lstm-and-gru-a4250bf6c090. acessado em 06/11/2019.

Chen, G., Li, J., and Wang, J. (2013). Evaluation of gene ontology semantic similarities on protein interaction datasets. *International Journal of Bioinformatics Research and Applications*, 9(2):173.

Chen, W.-H., Lu, G., Chen, X., Zhao, X.-M., and Bork, P. (2016). OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Research*, 45(D1):D940–D944.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Consortium, G. O. (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(90001):258D–261.

Cooper, G. M. and Hausman, R. E. (2007). *The Cell: A Molecular Approach*. ASM Press, 4 edition.

da Costa, W. L. O., de Aragão Araújo, C. L., Dias, L. M., de Sousa Pereira, L. C., Alves, J. T. C., Araújo, F. A., Folador, E. L., Henriques, I., Silva, A., and Folador, A. R. C. (2018). Functional annotation of hypothetical proteins from the exiguobacterium antarcticum strain b7 reveals proteins involved in adaptation to extreme environments, including high arsenic resistance. *PLOS ONE*, 13(6):e0198965.

da Silva, F. A. B., Carels, N., and Junior, F. P. S., editors (2018). *Theoretical and Applied Aspects of Systems Biology*. Springer International Publishing.

DeepAI (2020). *What is an Epoch?* https://deepai.org/machine-learning-glossary-and-terms/epoch. acessado em 08/02/2020.

der Maaten, L. V. (2020). *t-SNE*. https://lvdmaaten.github.io/tsne/. acessado em 22/02/2020.

du Plessis, L., Skunca, N., and Dessimoz, C. (2011). The what, where, how and why of gene ontology–a primer for bioinformaticians. *Briefings in Bioinformatics*, 12(6):723–735.

DZIF German Center for Infection Research (2019). *New inhibitor for persistent bacterial biofilms*. https://www.dzif.de/en/new-inhibitor-persistent-bacterial-biofilms. acessado em 05/11/2019.

Edwards, D., Stajich, J., and Hansen, D. (2009). *Bioinformatics: Tools and Applications*. Biomedical and Life Sciences. Springer New York.

EMBL-EBI (2019). *Running InterProScan 5*. https://github.com/ebi-pf-team/interproscan/wiki/HowToRun. acessado em 08/11/2019.

Emms, D. M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1).

Fundação Oswaldo Cruz (2019). *Computational Modeling of Multidrug-resistant Bacteria*. http://pseudomonas.procc.fiocruz.br. acessado em 10/10/2019.

Furuno, M. (2003). Cds annotation in full-length cdna sequence. *Genome Research*, 13(6):1478–1487.

Gardner, P. P., Paterson, J. M., Ashari-Ghomi, F., Umu, S. U., McGimpsey, S., and Pawlik, A. (2016). A meta-analysis of bioinformatics software benchmarks reveals that publication-bias unduly influences software accuracy.

Garrels, J. (2001). Proteome. In *Encyclopedia of Genetics*, pages 1575–1578. Elsevier.

Gene Ontology Consortium (2016). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338.

Gene Ontology Consortium (2019). *Gene Ontology Overview*. http://geneontology.org/docs/ontology-documentation/. acessado em 01/11/2019.

Georgevici, A. I. and Terblanche, M. (2019). Neural networks and deep learning: a brief introduction. *Intensive Care Medicine*, 45(5):712–714.

Goldberg, A. P., Szigeti, B., Chew, Y. H., Sekar, J. A., Roth, Y. D., and Karr, J. R. (2018). Emerging whole-cell modeling principles and methods. *Current Opinion in Biotechnology*, 51:97–102.

Goldman, A. D. and Landweber, L. F. (2016). What is a genome? *PLOS Genetics*, 12(7):e1006181.

Google (2019). *Google Colab*. https://colab.research.google.com. acessado em 01/11/2019.

Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talon, M., Dopazo, J., and Conesa, A. (2008). High-throughput functional annotation and data mining with the blast2go suite. *Nucleic Acids Research*, 36(10):3420–3435.

Gruber, A., Durham, A., Huynh, C., and et al. (2008). *Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach [Internet]*. National Center for Biotechnology Information (US), Bethesda (MD).

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Gurney, K. (2014). *An Introduction to Neural Networks*. CRC Press.

Haykin, S. (2009). *Neural networks and learning machines*. Prentice-Hall, 3 edition.

Henry, V. J., Bandrowski, A. E., Pepin, A.-S., Gonzalez, B. J., and Desfeux, A. (2014). OMICtools: an informative directory for multi-omic data analysis. *Database*, 2014(0):bau069–bau069.

Hill, S. T., Kuintzle, R., Teegarden, A., Merrill, E., Danaee, P., and Hendrix, D. A. (2018). A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. *Nucleic Acids Research*, 46(16):8105–8113.

Hu, J. C., Karp, P. D., Keseler, I. M., Krummenacker, M., and Siegele, D. A. (2009). What we can learn about escherichia coli through application of gene ontology. *Trends in Microbiology*, 17(7):269–278.

IEEE Spectrum (2018). *The 2018 Top Programming Languages*. https://spectrum.ieee.org/at-work/innovation/the-2018-top-programming-languages. acessado em 07/12/2019.

Ijaq, J., Malik, G., Kumar, A., Das, P. S., Meena, N., Bethi, N., Sundararajan, V. S., and Suravajhala, P. (2019). A model to predict the function of hypothetical proteins through a nine-point classification scoring schema. *BMC Bioinformatics*, 20(1).

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., and Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.

Kaji, D. A., Zech, J. R., Kim, J. S., Cho, S. K., Dangayach, N. S., Costa, A. B., and Oermann, E. K. (2019). An attention based deep learning model of clinical events in the intensive care unit. *PLOS ONE*, 14(2):e0211057.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2015). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462.

Kanter, A. S., Borland, R., Barasa, M., Iiams-Hauser, C., Velez, O., Kaonga, N. N., and Berg, M. (2012). The importance of using open source technologies and common standards for interoperability within eHealth: Perspectives from the millennium villages project. In *Advances in Health Care Management*, pages 189–204. Emerald Group Publishing Limited.

Karp, P. D., Weaver, D., Paley, S., Fulcher, C., Kubo, A., Kothari, A., Krummenacker, M., Subhraveti, P., Weerasinghe, D., and Gama-Castro, S. e. a. (2014). The ecocyc database. *EcoSal Plus*, 6(1).

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., and Duran, C. e. a. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649.

Keseler, I. M., Bonavides-Martinez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., Krummenacker, M., Nolan, L. M., Paley, S., and Paulsen, I. T. e. a. (2009). Ecocyc: A comprehensive view of escherichia coli biology. *Nucleic Acids Research*, 37(Database):D464–D470.

Kim, Y. J., Oh, Y., Park, S., Cho, S., and Park, H. (2013). Stratified sampling design based on data mining. *Healthcare Informatics Research*, 19(3):186.

Klockgether, J., Munder, A., Neugebauer, J., Davenport, C. F., Stanke, F., Larbig, K. D., Heeb, S., Schock, U., Pohl, T. M., Wiehlmann, L., and Tummler, B. (2009). Genome diversity of pseudomonas aeruginosa PAO1 laboratory strains. *Journal of Bacteriology*, 192(4):1113–1121.

Klockgether, J. and Tümmler, B. (2017). Recent advances in understanding pseudomonas aeruginosa as a pathogen. *F1000Research*, 6:1261.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, 39(1):309–338.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., and Wassenaar, T. e. a. (2015). Insights from 20 years of bacterial genome sequencing. *Functional Integrative Genomics*, 15(2):141–161.

Lapatas, V., Stefanidakis, M., Jimenez, R. C., Via, A., and Schneider, M. V. (2015). Data integration in biological research: an overview. *Journal of Biological Research-Thessaloniki*, 22(1).

Lei, X. and Yang, X. (2018). A new method for predicting essential proteins based on participation degree in protein complex and subgraph density. *PLOS ONE*, 13(6):e0198998.

Li, H., Gong, X.-J., Yu, H., and Zhou, C. (2018). Deep neural network based predictions of protein interactions using primary sequences. *Molecules*, 23(8):1923.

Lipton, Z. C., Elkan, C., and Naryanaswamy, B. (2014). Optimal thresholding of classifiers to maximize f1 measure. In *Machine Learning and Knowledge Discovery in Databases*, pages 225–239. Springer Berlin Heidelberg.

Long, H., Liao, B., Xu, X., and Yang, J. (2018). A hybrid deep learning model for predicting protein hydroxylation sites. *International Journal of Molecular Sciences*, 19(9):2817.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2002). Semantic similarity measures as tools for exploring the gene ontology. In *Biocomputing 2003*. World Scientific.

Machine Intellegence (2020). *LSTM-Long Short Term Memory*. http://www.machineintellegence.com/lstm-long-short-term-memory/. acessado em 08/02/2020.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, pages 276–282.

Metwally, A. A., Yu, P. S., Reiman, D., Dai, Y., Finn, P. W., and Perkins, D. L. (2019). Utilizing longitudinal microbiome taxonomic profiles to predict food allergy via long short-term memory networks. *PLOS Computational Biology*, 15(2):e1006693.

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H.-Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Nuka, G., Orengo, C., Pandurangan, A. P., Paysan-Lafosse, T., Pesseat, S., Potter, S. C., Qureshi, M. A., Rawlings, N. D., Redaschi, N., Richardson, L. J., Rivoire, C., Salazar, G. A., Sangrador-Vegas, A., Sigrist, C. J. A., Sillitoe, I., Sutton, G. G., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Yong, S.-Y., and Finn, R. D. (2018). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1):D351–D360.

Mohamed Elfil, A. N. (2019). *Illustrated Guide to LSTM's and GRU's: A step by step explanation*. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21. acessado em 27/10/2019.

MySQL (2020). *Limits on Number of Databases and Tables*. https://dev.mysql.com/doc/mysql-reslimits-excerpt/5.6/en/database-count-limit.html. acessado em 08/02/2020.

National Center for Biotechnology Information (US) (2019). *PubMed Help [Internet]*. https://www.ncbi.nlm.nih.gov/books/NBK3827/. acessado em 01/11/2019.

NCBI National Center for Biotechnology Information (2018a). *Genbank sample record*. https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html. acessado em 06/05/2018.

NCBI National Center for Biotechnology Information (2018b). *GENOME REPORTS - Prokaryotes.txt*. ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt. acessado em 26/03/2018.

NCBI National Center for Biotechnology Information (2018c). *Proper Use of locus tag in Genome Submissions*. https://www.ncbi.nlm.nih.gov/genomes/locustag/Proposal.pdf. acessado em 11/05/2018.

NCBI National Center for Biotechnology Information (2019). *What is genome annotation* https://support.nlm.nih.gov/knowledgebase/article/KA-03574/en-us. acessado em 03/11/2019.

Nichio, B. T. L., Marchaukoski, J. N., and Raittz, R. T. (2017). New tools in orthology analysis: A brief review of promising perspectives. *Frontiers in Genetics*, 8.

Pando-Robles, R. V., Lanz-Mendoza, H., RV, P.-R., and H, L.-M. (2009). La importancia de la proteómica en la salud pública: The significance of proteomics in public health. *Salud Pública de México*, 51.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2012). Scikit-learn: Machine learning in python. *CoRR*, abs/1201.0490.

Peng, J., Zhang, X., Hui, W., Lu, J., Li, Q., Liu, S., and Shang, X. (2018). Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach. *BMC Systems Biology*, 12(S2).

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7):e1000443.

Qin, C., Sun, Y., and Dong, Y. (2017). A new computational strategy for identifying essential proteins based on network topological properties and biological information. *PLOS ONE*, 12(7):e0182031.

Reeves, G. A., Talavera, D., and Thornton, J. M. (2009). Genome and proteome annotation: organization, interpretation and integration. *Journal of The Royal Society Interface*, 6(31):129–147.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

Retson, T. A., Besser, A. H., Sall, S., Golden, D., and Hsiao, A. (2019). Machine learning and deep neural networks in thoracic and cardiovascular imaging. *Journal of Thoracic Imaging*, 34(3):192–201.

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.8.3.

Rigden, D. J. and Fernández, X. M. (2018). The 26th annual nucleic acids research database issue and molecular biology database collection. *Nucleic Acids Research*, 47(D1):D1–D7.

Riggio, C. (2019). *What's the deal with Accuracy, Precision, Recall and F1?* https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021. acessado em 24/11/2019.

RStudio, Inc. (2018). *RStudio - Open source and Enterprise-ready professional software for R*. https://www.rstudio.com/products/rstudio/. acessado em 26/03/2018.

SAS (2019). *Redes neurais - o que são e qual sua importância*. https://www.sas.com/pt_br/insights/analytics/neural-networks.html. acessado em 27/10/2019.

Scalegrid.io (2020). *2019 Database Trends – SQL vs. NoSQL, Top Databases, Single vs. Multiple Database Use*. https://scalegrid.io/blog/2019-database-trends-sql-vs-nosql-top-databases-single-vs-multiple-database-use/. acessado em 29/01/2020.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLOS Biology*, 14(8):e1002533.

Sheehan, B., Quigley, A., Gaudin, B., and Dobson, S. (2008). A relation based measure of semantic similarity for gene ontology annotations. *BMC Bioinformatics*, 9(1).

Sidey-Gibbons, J. A. M. and Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19(1).

Silberschatz, A., Korth, H. F., and Sudarshan, S. (2001). *Database system concepts*. McGraw-Hill Education, 7 edition. páginas 1,244.

Solid IT (2019). *Knowledge Base of Relational and NoSQL Database Management Systems*. https://db-engines.com/en/ranking. acessado em 07/12/2019.

Stack Overflow (2020). *Developer Survey Results 2019*. https://insights.stackoverflow.com/survey/2019/technology. acessado em 08/02/2020.

Stanford University (2019). *CS231n Convulutional Neural Networks for Visual Recognition*. http://cs231n.github.io/neural-networks-1/. acessado em 24/11/2019.

Tacconelli, E., Carrara, E., Savoldi, A., Harbarth, S., Mendelson, M., Monnet, D. L., Pulcini, C., Kahlmeter, G., Kluytmans, J., Carmeli, Y., Ouellette, M., Outterson, K., Patel, J., Cavaleri, M., Cox, E. M., Houchens, C. R., Grayson, M. L., Hansen, P., Singh, N., Theuretzbacher, U., Magrini, N., Aboderin, A. O., Al-Abri, S. S., Jalil, N. A., Benzonana, N., Bhattacharya, S., Brink, A. J., Burkert, F. R., Cars, O., Cornaglia, G., Dyar, O. J., Friedrich, A. W., Gales, A. C., Gandra, S., Giske, C. G., Goff, D. A., Goossens, H., Gottlieb, T., Blanco, M. G., Hryniewicz, W., Kattula, D., Jinks, T., Kanj, S. S., Kerr, L., Kieny, M.-P., Kim, Y. S., Kozlov, R. S., Labarca, J., Laxminarayan, R., Leder, K., Leibovici, L., Levy-Hara, G., Littman, J., Malhotra-Kumar, S., Manchanda, V., Moja, L., Ndoye, B., Pan, A., Paterson, D. L., Paul, M., Qiu, H., Ramon-Pardo, P., Rodríguez-Baño, J., Sanguinetti, M., Sengupta, S., Sharland, M., Si-Mehand, M., Silver, L. L., Song, W., Steinbakk, M., Thomsen, J., Thwaites, G. E., van der Meer, J. W., Kinh, N. V., Vega, S., Villegas, M. V., Wechsler-Fördös, A., Wertheim, H. F. L., Wesangula, E., Woodford, N., Yilmaz, F. O., and Zorzet, A. (2018). Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *The Lancet Infectious Diseases*, 18(3):318–327.

The Gene Ontology Consortium (2018). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338.

The UniProt Consortium (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 46(5):2699–2699.

Torres. Ai, J. (2018). *Neural Networks*. https://towardsdatascience.com/basic-concepts-of-neural-networks-1a18a7aa2bd2. acessado em 10/10/2019.

Towards Data Science (2020). *LSTM — nuggets for practical applications*. https://towardsdatascience.com/lstm-nuggets-for-practical-applications-5beef5252092. acessado em 31/01/2020.

Tutar, Y. (2012). Pseudogenes. *Comparative and Functional Genomics*, 2012:1–4.

Vallet-Gely, I. and Boccard, F. (2013). Chromosomal organization and segregation in pseudomonas aeruginosa. *PLoS Genetics*, 9(5):e1003492.

Venkova, T., Yeo, C. C., and Espinosa, M. (2018). Editorial: The good, the bad, and the ugly: Multiple roles of bacteria in human life. *Frontiers in Microbiology*, 9.

W3C World Wide Web Consortium (2018). *The World Wide Web Consortium*. https://www.w3.org/standards/semanticweb/ontology. acessado em 07/02/2018.

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281.

Wang, Z., Chen, Y., and Li, Y. (2004). A brief review of computational gene prediction methods. *Genomics, Proteomics & Bioinformatics*, 2(4):216–221.

Wanichthanarak, K., Fahrmann, J. F., and Grapov, D. (2015). Genomic, proteomic, and metabolomic data integration strategies. *Biomarker Insights*, 10s4:BMI.S29511.

Watson, P. and Petrie, A. (2010). Method agreement analysis: A review of correct methodology. *Theriogenology*, 73(9):1167–1179.

Winsor, G. L. (2004). Pseudomonas aeruginosa genome database and pseudocap: facilitating community-based, continually updated, genome annotation. *Nucleic Acids Research*, 33(Database issue):D338–D343.

Winsor, G. L., Griffiths, E. J., Lo, R., Dhillon, B. K., Shay, J. A., and Brinkman, F. S. L. (2015). Enhanced annotations and features for comparing thousands ofPseudomonasgenomes in the pseudomonas genome database. *Nucleic Acids Research*, 44(D1):D646–D653.

World Health Organization (2019). *Patient Safety*. https://www.who.int/news-room/fact-sheets/detail/patient-safety. acessado em 10/10/2019.

Zhang, Z. (2016). A gentle introduction to artificial neural networks. *Annals of Translational Medicine*, 4(19):370–370.

Zhao, C. and Wang, Z. (2018). GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. *Scientific Reports*, 8(1).

Zhulin, I. B. (2015). Databases for microbiologists: TABLE 1. *Journal of Bacteriology*, 197(15):2458–2467.