UNIVERSIDADE FEDERAL DO PARANÁ

LETÍCIA GRAZIELA COSTA SANTOS DE MATTOS

ANÁLISE E CARACTERIZAÇÃO DE GRANDES GRUPOS DE PROTEÍNAS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

CURITIBA
2019

LETÍCIA GRAZIELA COSTA SANTOS DE MATTOS

ANÁLISE E CARACTERIZAÇÃO DE GRANDES GRUPOS DE PROTEÍNAS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica da Universidade Federal do Paraná, como parte das exigências para a obtenção do título de Mestre em Bioinformática.

Orientador: Prof. Dr. Roberto Tadeu Raittz

CURITIBA

2019

Catalogação na publicação
Sistema de Bibliotecas UFPR
Biblioteca de Educação Profissional e Tecnológica

Mattos, Leticia Graziela Costa Santos de
M444 Análise e caracterização de grandes grupos de proteínas utilizando mineração de dados / Leticia Graziela Costa Santos de Mattos. - Curitiba, 2019.
102 p.: il.

Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Curso de Pós-Graduação em Bioinformática, 2019.
Orientador: Roberto Tadeu Raitz

1. Mineração de dados (Computação). 2. Análise por conglomerados. 3. Bioinformática. I. Raitz, Roberto Tadeu. II. Título. III. Universidade Federal do Paraná.

CDD 006.312

Elaboração: Angela Pereira de Farias Mengatto - CRB 9/1002

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA

Pós-Graduação em Bioinformática WWW.BIOINFO.UFPR.BR
E-mail: bioinfo@ufpr.br **Tel: 41 33614906**

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em BIOINFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **LETÍCIA GRAZIELA COSTA SANTOS DE MATTOS** intitulada: **"Análise e Caracterização de Grandes Grupos de Proteínas Utilizando Técnicas de Mineração de Dados"**, após terem inquirido a aluna e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa. A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 5 de fevereiro de 2019.

Dr. Roberto Tadeu Raittz
Presidente/Programa de Pós-graduação em Bioinformática – UFPR

Drª. Ana Claudia Bonatto
Avaliadora Externa/Programa de Pós-graduação em Genética - UFPR

Dr. Fábio Passetti
Avaliador Interno/Programa de Pós-graduação em
Bioinformática – UFPR/Instituto Carlos Chagas-FIOCRUZ

Dedico à minha inesquecível bisavó
Andreza Theodoro de Moura (*in memoriam*)
e ao meu esposo **Vitor Hugo de Mattos**,
por sempre enxergarem, em meio aos meus muitos defeitos,
as minhas melhores qualidades.

AGRADECIMENTOS

A elaboração desta dissertação, assim como todo o percurso acadêmico que a antecedeu, só foi possível graças a contribuição, carinho e apoio de várias pessoas, às quais gostaria de exprimir algumas palavras de agradecimento e reconhecimento, sob a pena de me esquecer de alguém.

Em primeiro lugar agradeço a Deus por me mostrar que os Seus planos para a minha vida são muito maiores do que aquilo que eu pedi ou imaginei. Por me inspirar e capacitar.

Ao meu esposo Vitor Hugo de Mattos por todo amor, carinho, paciência, apoio e compreensão que dedicou a mim nos últimos anos. Eu amo você.

Aos meus pais biológicos, Vanderléia Aparecida Batista da Costa e Pautilho Alberto dos Santos, e também aos meus pais adotivos Mauri Cesar Teixeira e Ana Christina Jamnik Teixeira por lutar constantemente para que os meus sonhos se tornem realidade, me transmitindo os mais valiosos saberes. Amo vocês.

Aos meus irmãos em Cristo por todo amor e apoio emocional nos momentos difíceis.

Aos meus amigos do Laboratório de Bioinformática, em especial à Camilla Reginatto de Pierri, Antônio Camilo da Silva, Aryel Marlus Répula de Oliveira, Diogo de Jesus Soares Machado, Camila Pereira Perico e Bruno Thiago de Lima Nichio. Também aos nossos brilhantes estagiários e alunos de Iniciação Científica Jéssica Maria Magno, Monique Schreiner e Danrley. Vocês foram essenciais no desenvolvimento deste trabalho.

Ao meu querido e admirado orientador Prof. Dr. Roberto Tadeu Raittz pela paciência, amizade e conselhos.

A Profa. Dra. Ana Claudia Bonatto e ao Dr. Fabio Passetti por atenderem ao convite para compor minha banca avaliadora e pela disponibilidade em participar da correção desta dissertação.

A todos os professores, alunos e funcionários do Programa de Pós-Graduação em Bioinformática da UFPR por me receberem com tanto carinho, compartilhando comigo uma pequena porção do que vocês sabem.

Agradeço especialmente a Suzanna Gobetti, nossa amada e queria secretária pela amizade e apoio nestes dois anos.

A CAPES pelo apoio financeiro.

A todos os que direta ou indiretamente me ajudaram e torceram por mim, os meus sinceros agradecimentos.

“Porque d’Ele, por meio d’Ele e para Ele são todas as coisas.
A Ele seja a Glória eternamente. “Romanos 11:36

“Questions of science, science and progress
Do not speak as loud as my heart”
Coldplay

RESUMO

Nas últimas décadas com o rápido desenvolvimento de disciplinas como a genômica e a proteômica, a quantidade de informação biológica que é produzida e armazenada diariamente nos Bancos de Dados de proteínas tem aumentado de forma rápida e irregular, tornando a aplicação e o desenvolvimento de técnicas de mineração de dados cada vez mais importante. No caso dos bancos de dados de sequências biológicas, problemas na qualidade dos dados como alto nível de redundância e artefatos de anotação tornaram as técnicas de clusterização uma das formas mais rápidas e eficientes de solucionar problemas como armazenamento, curadoria e busca contra os bancos de dados. Entretanto, analisando criteriosamente o estado da arte na clusterização de bancos de dados de sequências biológicas percebe-se a necessidade de reprocessar os resultados quando se obtêm clusters muito grandes se comparado à média do banco. Assim, neste contexto, este trabalho propôs a criação de um pipeline para a aplicação de técnicas de mineração de dados com o objetivo de caracterizar grandes conjuntos de dados gerados após a clusterização de bancos de dados de sequências biológicas. Análises realizadas com base em um estudo de caso biológico permitiram a criação de um pipeline baseado em inferência de homologia, anotações funcionais de Gene Ontology e técnicas de mineração de texto desenvolvidas neste trabalho. Os resultados mostram que, de acordo com a consistência da anotação da função intracluster, os maiores clusters requerem reprocessamento quando o banco de dados foi clusterizado com o valor de corte de 50% de identidade. O algoritmo de clusterização de texto desenvolvido para o pipeline foi preciso e eficiente para reclusterizar os conjuntos de dados utilizados neste trabalho. Os resultados deste trabalho levam a recomendações práticas para usos mais eficazes dos resultados das ferramentas de clusterização de sequências biológicas.

Palavras-chave: Clusterização. Bancos de Dados Biológicos. Mineração de Dados.

ABSTRACT

In the last years, the rapid development of disciplines such as genomics and proteomics generated an amount of biological information that is daily stored in protein databases. Thus, these biological databases have increased rapidly and irregularly, making primordial the application and development of data mining techniques. In the case of biological sequence databases, data quality problems such as high level of redundancy and annotation artifacts have made clustering techniques one of the fastest and most efficient ways of solving problems such as storage, curation and database search. However, by carefully analyzing the State of the Art in clustering of biological sequence databases, it's noticed that's necessary to reprocess the results when very large clusters are obtained, but the best way to do this reprocessing is yet an open question. Thus, in this context, this work proposed the creation of a pipeline for the application of data mining techniques with the aim of characterizing large proteins datasets generated after clustering biological sequence databases. Analyzes carried out based on a biological case study allowed the creation of a pipeline based on homology inference, functional annotations of Gene Ontology and text mining techniques developed in this work. Results show that according to intracluster function annotation consistency, clusters with large size require reprocessing when the database was clustered with self-score of 50% of identity. The text clustering algorithm developed for the pipeline was accurate and efficient in reclustering the datasets. This evaluation leads to practical recommendations for more effective uses of the sequence clustering tools results.

Keywords: Clustering. Biological Databases. Data Mining.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BIND – the Biomolecular Interaction Network Database

CSNDB – Cell-signaling networks database

DIP – Database of Interacting Proteins

DNA -Ácido Desoxirribonucleico

DSSP – Define Secondary Structure of Proteins

GO – Gene Ontology

HMM – Hidden Markov Model

IA – Inteligência Artificial

KDD – Knowledge Discovery Databases

KEGG – Kyoto Encyclopedia of Genes and Genomes

MD – Mineração de Dados

MINT – The Molecular INTeracting Proteins

MLP – Rede Neuronal de Multipla Camada

NR – *Non-redundant Data Sequences of Proteins*

PDB – Protein Data Bank

PIR – Protein Information Resource

RefSeq – NCBI Reference Sequence Database

RNA – Rede Neuronal Artificial

RNAt – Ácido Ribonucleico Transportador

SCOP – Structural Classification of Proteins

SMART – Simple Modular Architeture Tool

SPAD – Signaling Pathway Database

WIT – What Is There

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, o rápido desenvolvimento de disciplinas como a genômica e a proteômica tem gerado uma grande quantidade de dados biológicos que é armazenada diariamente nos bancos de dados. Este aumento rápido e irregular tornou a aplicação e o desenvolvimento de técnicas de mineração de dados cada vez mais importante em estudos de genomas e proteomas. A análise de grandes conjuntos de dados biológicos requer extrair dos dados informações que façam sentido, por meio de inferência ou generalizando os dados (CHAWLA; SHARMA, 2016; EMMS; STEVEN, 2015; KELIL et l, 2007; LI; JAROSZEWSKI; GODZIK, 2001).

A busca de similaridade entre sequências biológicas é uma ferramenta poderosa para predizer a função de uma proteína desconhecida, pois informações funcionais podem ser transferidas entre proteínas homólogas. Sequências de aminoácidos com similaridades interessantes e inesperadas entre si podem ocorrer em todo um intervalo de 0 a 100% de identidade, entretanto os bancos de dados têm uma superabundância de sequências que apresentam cerca de 90% de identidade (HOLM & SANDER, 1998).

O banco de dados *Non-redundant Data Sequences of Proteins (NR),* que foi compilado pelo NCBI a partir de diferentes bancos de dados de proteínas utilizando o algoritmo *nrdb*, tinha como objetivo remover todas as entradas idênticas para diminuir sua redundância, entretanto, ainda assim este é um banco altamente redundante. Assim, no banco de dados NR uma família bem caracterizada de proteínas pode conter inúmeras entradas idênticas ou quase idênticas de algumas espécies, mas poucos homólogos de outras fontes (LI; JAROSZEWSKI; GODZIK, 2002).

Uma abordagem para solucionar este problema é utilizar técnicas de Mineração de Dados (*Data Mining*) como a clusterização, pois ela diminui o tempo de busca contra o banco e simplifica a organização dos resultados, além do fato de que as sequências em um cluster tipicamente tem uma função biológica em comum (HOLM & SANDER, 1998; LI; JAROSZEWSKI; GODZIK, 2002).

Entretanto, ao observar o Estado da Arte na clusterização de bancos de dados de sequências biológicas de forma mais criteriosa, alguns autores têm observado que em alguns casos pode existir a necessidade de reprocessar os

resultados ao invés de utilizá-los em sua forma bruta (CHEN, et. al., 2018). Um exemplo levantado por Chen e colaboradores são os maiores clusters gerados após a aplicação de métodos como CD-HIT (LI; JAROSZEWSKI; GODZIK, 2001; FU, et. al., 2012) e UCLUST (EDGAR, 2010).

Sendo assim, a atual pesquisa teve como objetivo principal aplicar técnicas de mineração de dados na análise de grandes grupos de proteínas e criar um pipeline para interpretação dos resultados obtidos na clusterização de grandes conjuntos de dados de sequências biológicas.

2 REFERENCIAL TEÓRICO

2.1 Mineração de Dados

A mineração de dados (MD), também conhecida como *data mining* ou *Knowledge Discovery Databases* (KDD), é a área que busca novos padrões e relacionamentos interessantes em uma grande quantidade de dados. A MD é definida como o processo de descoberta de novas correlações significativas, padrões e tendências, trabalhando em quantidades abundantes de dados armazenados em depósitos (CHAWLA; SHARMA, 2016; GONZALEZ, et al, 2016; MASOOD; KHAN, 2015).

No trabalho de Fayyad et al. (1996), o processo de descoberta de conhecimento é descrito em diferentes etapas, começando na seleção de dados, pré-processamento, transformação de dados, mineração de dados e interpretação.

Historicamente, o KDD foi construído em três campos: Aprendizado de Máquina, Bancos de Dados e Inteligência Artificial, com o intuito de projetar e desenvolver ferramentas e estruturas de suporte que permitam ao usuário final ganhar insights na natureza dos dados que são excepcionalmente grandes (HOLZINGER; DEHMER; JURISICA, 2014).

Os dois primeiros objetivos da mineração de dados são descrever e predizer, sendo que as principais tarefas que envolvem a mineração de dados para encontrar novos padrões significativos com base nos dados, são (CASTANHEIRA, 2008; CHAWLA; SHARMA, 2016; GONZALEZ, et al, 2016):

- **Classificação**: A Classificação é uma função de aprendizado que mapeia um item de dados em classes predefinidas.
- **Estimação**: Usada para definir um valor para uma variável contínua desconhecida, como por exemplo, a altura de uma pessoa.
- **Predição**: É similar a classificação e a estimação, exceto no fato de que os dados são classificados de acordo com um valor estimado.
- **Regras de Associação ou Análises de Associação**: Determina quais objetos devem permanecer unidos.

- **Clusterização**: É uma classificação não supervisionada de padrões como observações, objetos ou vetores de características em um número de subgrupos ou clusters.
- **Descrição e visualização**: Representa os dados utilizando técnicas de visualização.

2.1.2 Mineração de Textos

A mineração de textos ou *text mining*, é um subcampo da mineração de dados que procura extrair informações novas a partir de fontes não estruturadas ou semiestruturadas (GONZALEZ, et al, 2016), podendo ser entendida como o estudo e a prática de extrair tendências, regras e padrões com base em textos completos de artigos científicos digitais, utilizando métodos analíticos e fundamentos da linguística computacional. Assim, dado um conjunto de documentos, os métodos de mineração de textos buscam automaticamente novos padrões e relacionamentos entre estes documentos (COHEN; HUNTER, 2008; FAIIAZE et al., 2012; WOSZEZENKI; GONÇALVES, 2013; SULLIVAN, 2001).

O rápido progresso das pesquisas no domínio biomédico causou um exorbitante aumento no número de publicações científicas e com isso técnicas de mineração de textos começaram a ser utilizadas desde a década de 90, com o objetivo de automatizar a extração de informações importantes presentes nos textos biomédicos. Assim, o domínio biomédico é considerado uma das áreas mais interessantes para a aplicação da mineração de texto, devido ao potencial impacto das informações que podem ser descobertas e do volume de informação disponível.

Na área biomédica, as aplicações de mineração de textos geralmente procuram extrair padrões como relações proteínas-proteína, gene-proteína, droga-proteína, gene-gene e gene-doença, para descobrir novos tratamentos, diagnósticos e prevenções (WOSZEZENKI; GONÇALVES, 2013; COHEN; HERSH, 2005).

Hoje, a principal fonte utilizada para a descoberta de conhecimento na biomedicina é a literatura científica armazenada no PubMed. O PubMed disponibiliza mais de 28 milhões de citações e resumos da literatura biomédica do MEDLINE, jornais de ciências da vida e livros *online* das áreas da medicina, enfermagem, odontologia, medicina veterinária, biologia, bioquímica, evolução molecular e muitas outras (NCBI, 2017; WOSZEZENKI; GONÇALVES, 2013).

Existem três tipos principais de abordagens de mineração de textos que tem sido empregadas no domínio biomédico: os métodos baseados em coocorrência, que procuram conceitos que ocorrem na mesma unidade, e outras duas metodologias mais comuns e sofisticadas, que são as abordagens baseadas em regras e conhecimento e as que se baseiam em aprendizado de máquina (COHEN; HUNTER, 2008).

2.1.3 Aplicações da Mineração de Dados em Bioinformática

A Bioinformática e a Mineração de Dados têm se desenvolvido rapidamente como ciências interdisciplinares. As abordagens de mineração de dados são ideais para a Bioinformática, pois esta é uma área rica em dados, mas que carece de uma teoria abrangente sobre a organização da vida a nível molecular (HOLZINGER; DEHMER; JURISICA, 2014).

O problema nas ciências da vida está no fato de que os modelos de dados biomédicos são altamente complexos, tornando a análise manual pelos usuários finais da informação difícil e muitas vezes impossível. Assim, minerar bancos de dados biológicos e de outras ciências da vida relacionadas como medicina e neurociência, auxilia na extração de informações úteis a partir de densos conjuntos de dados (HOLZINGER; DEHMER; JURISICA, 2014).

Hoje, algumas das aplicações da mineração de dados na bioinformática são a limpeza de dados, previsão de localização subcelular de proteínas, reconstrução de redes de interação entre proteínas e genes, descoberta de novos genes, diagnóstico e prognóstico de doenças, inferência de novas funções proteicas e detecção de domínios conservados (CHAWLA; SHARMA, 2016). Outra técnica de MD amplamente utilizada pela Bioinformática é a mineração de textos a partir de bases textuais biomédicas (WOSZEZENKI; GONÇALVES, 2013).

Entretanto, o uso efetivo de técnicas de mineração de dados é muitas vezes prejudicado por algumas características dos bancos de dados biológicos, como tamanho, número, falta de ontologia padrão, diversidade, heterogeneidade dos dados, integração entre os bancos, inconsistência e alta dimensionalidade dos dados. Assim, o desafio não é apenas extrair informação significativa a partir destes dados, mas obter conhecimento para descobrir e compreender algo que ainda não foi visto previamente e trazer sentido aos dados (CHAWLA; SHARMA, 2016;

HOLZINGER; DEHMER; JURISICA, 2014).

2.2 Bancos de Dados Biológicos

Bancos de dados são sistemas utilizados para armazenar e recuperar qualquer tipo de dado (ULLMAN; WIDOM, 1997). Bancos de dados biológicos são bancos que armazenam informações oriundas de análises de alto rendimento, análises computacionais de áreas como Genômica, Proteômica, Filogenômica e também da literatura biomédica (EMBL; EBI, 2018; SARAVANAN; DEVI, 2012).

O planejamento, desenvolvimento e manutenção de um banco de dados biológicos é uma das principais áreas da Bioinformática. Normalmente os bancos de dados biológicos são classificados como primário ou secundário (EMBL; EBI, 2018; SARAVANAN; DEVI, 2012; ZOU, et al., 2015). Os bancos de dados primários são povoados com dados gerados em análises experimentais, como por exemplo sequências de aminoácidos, nucleotídeos e estruturas tridimensionais de proteínas. Os bancos de dados secundários contêm dados dos resultados obtidos na análise dos dados primários. Este tipo de banco de dados costuma ser altamente curado, pois utiliza a combinação de algoritmos computacionais e análise manual (EMBL; EBI, 2018).

2.2.1 Bancos de dados de proteínas

Um banco de dados de proteínas pode ser definido como uma coleção de dados que foi construída a partir de informações biológicas, físicas e químicas que podem incluir sequências de aminoácidos, estruturas e informações conformacionais de proteínas, interações proteína-proteína, e outras características como sítios ativos e funções moleculares. Alguns bancos de dados de proteínas são compilados a partir da tradução de sequências de DNA de diferentes bancos de genomas, sendo um importante recurso devido ao fato de que as proteínas são as responsáveis pela maioria das funções biológicas (KWON; CHO; PAIK, 2006; NATURE, 2017; NCBI, 2017).

Os bancos de dados de proteínas podem ser divididos em dois tipos principais:

I. Universal: contém proteínas de diversas espécies conhecidas que já foram caracterizadas;
II. Especializado: abrange as proteínas de um grupo específico ou famílias de proteínas de determinadas espécies;

É importante ressaltar que os bancos de dados de proteínas podem ser classificados em categorias mais específicas que estas de acordo com o tipo de informação que se deseja obter (CHEN; HUANG; WU, 2017):

- Bancos de dados de sequências (Swiss-Prot, NR, GenBank, RefSeq, TrEMBL, PIR);
- Bancos de dados de estruturas (PDB, CATH, Dali, DSSP, SCOP, Swiss-MODEL);
- Bancos de dados de interação proteína-proteína (Bind, DIP, Mint);
- Bancos de dados de perfis, Famílias e Domínios (InterPro, PROSITE, Pfam, PRINTS, ProDom, SMART);
- Bancos de dados de 2D-PAGE (eletroforese bidimensional em gel poliacrilamida) (SWISS-2D PAGE, YPRC-PDS);
- Bancos de dados de vias metabólicas (ENZYME, KEGG, WIT, PathDB);
- Bancos de dados de vias de sinalização (TRANSPATH, CSNDB, SPAD);
- Bancos de dados de Química (ChEMBL);
- Bancos de dados de Expressão Gênica (Expression Atlas);
- Bancos de dados de Anotação de Genomas (Ensembl, Entrez Gene, UCSC);
- Bancos de dados de organismos específicos (FlyBase, MGD, neXtProt);
- Bancos de dados de Filogenômica (OMA);
- Bancos de dados de polimorfismo e mutação (dbSNP);
- Bancos de dados de Proteômica (PRIDE, PeptideAtlas);
- Bancos de dados de Ontologia (Gene Ontology).

2.2.2 Qualidade dos Dados nos Bancos de Dados Biológicos

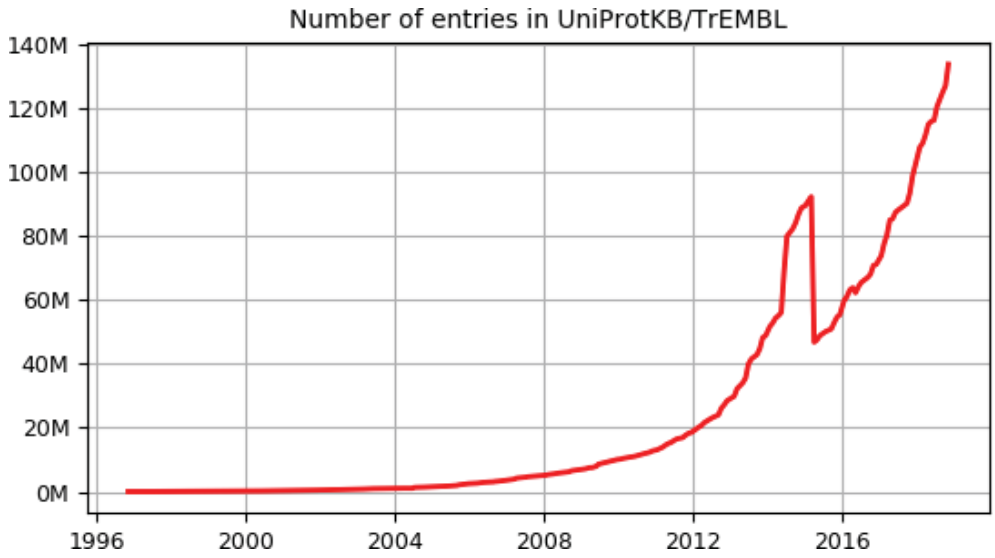
O rápido avanço de técnicas como o sequênciamento de alta vazão (REUTER; SPACEK; SNYDER, 2015) causou um crescimento estrondoso da

quantidade de dados biológicos submetidos nos bancos de dados, resultando em um alto nível de redundância e milhares de proteínas super-representadas (EMBL; SIB, 2016; CHEN, et al., 2018) em bancos como UniProtKB (FIGURA 1) (The UniProt Consortium, 2017) e NR (NCBI, 2018).

Redundância em um conjunto de dados de proteínas pode ser definida como a presença de proteínas muito similares (BULL; MULDOON; DOIG, 2013). Hoje a redundância é um obstáculo para o uso efetivo de um conjunto de dados de sequências biológicas por inúmeras razões, como por exemplo, o fato de que sequências com alto nível de identidade podem impedir a descoberta de novas relações entre proteínas ou tornar os dados tendenciosos. O alto nível de redundância em um banco de dados também causa outros problemas como lentidão nas buscas, além de tornar a análise dos resultados complexa e computacionalmente inviável fora de *workstations* (EMBL; SIB, 2016; BERNSTEIN, 2006).

Assim, com o incessante desenvolvimento de novas tecnologias de sequênciamento, encontrar novas formas de resolver estes problemas de qualidade e facilitar a interpretação destes dados pelos usuários se tornou essencial.

FIGURA 1 - NÚMERO DE ENTRADAS NO UNIPROTKB/TREMBL



FONTE: The UniProt Consortium, 2018.

LEGENDA: Resumo do crescimento do banco de dados UNIPROTKB/TrEMBL entre 1996 e 2018. Em sua última atualização (novembro de 2018) o banco continha 133.507.323 sequências.

Nos últimos anos, muitas soluções foram criadas para acabar com a redundância em bancos de dados biológicos. Dentre os softwares e metodologias utilizados para minimizar redundância estão UCLUST (EDGAR, 2010), SkipRedundant (RICE, LONGDEN, BLEASBY, 2000), CD-HIT (LI, 2006), Decreasy Redundancy (GASTEIGER, 2003) e PISCES (WANG, 2005). A diferença entre estes métodos está no tipo de alinhamento, que pode ser Global (NEEDLEMAN; WUNSCH, 1970) ou Local (SMITH; WATERMAN, 1981) e nos algoritmos de clusterização, sendo que o alinhamento local é mais sensível quando as sequências compartilham similaridade apenas em regiões específicas, ou quando a similaridade entre as sequências do banco é desconhecida (BULL, MULDOON, DOIG, 2013). O trabalho de Sikic & Carugo (2010) sugere que estes métodos são complementares e, portanto, o uso de mais de uma metodologia para remover redundância e interpretar os dados é recomendável.

Assim sendo, no caso de bancos de dados de sequências biológicas, uma solução para estes problemas de qualidade dos dados é clusterizar as sequências e então utilizar somente a sequência representante de cada grupo (LI; JAROSZEWSKI; GODZIK, 2002).

2.3 Clusterização de Bancos de Dados de Sequências Biológicas

Técnicas de clusterização têm sido amplamente utilizadas em estudos de genômica e proteômica. A clusterização é uma técnica não supervisionada de mineração de dados que constrói automaticamente conjuntos de objetos que têm características semelhantes de acordo com uma medida de distância, de forma que os dados em cada subconjunto tenham características mútuas (ARBELAITZ, et al., 2012). Este agrupamento ocorre de forma que o grau de associação entre os elementos do mesmo grupo é alto e entre elementos de grupos diferentes baixo (ARAUJO, 2007; GRONT; KOLINSKI, 2005; MASOOD; KHAN, 2015; RODRIGUES, et al., 2004).

A clusterização de dados é uma ferramenta chave em diferentes áreas da mineração de dados, principalmente quando se deseja interpretar grandes conjuntos de dados. As abordagens de clusterização de sequências biológicas podem ser classificadas em 3 categorias principais (CENTER FOR BIOINFORMATICS & COMPUTATIONAL BIOLOGY, 2010):

I. Aglomerativas: o algoritmo começa com cada sequência como um cluster independente e iterativamente une as sequências para formar clusters maiores.

II. Divisivas: o algoritmo começa com todas as sequências em um cluster e então iterativamente quebra o cluster em um conjunto de clusters menores.

III. Métodos de Partição: o algoritmo inicia com um conjunto pré-definido de clusters e então segue refinando os grupos.

Assim, a clusterização é um passo fundamental na análise de dados, pois ela é capaz de identificar grupos relacionados que podem ser utilizados como o ponto de partida para explorar outras associações (CENTER FOR BIOINFORMATICS & COMPUTATIONAL BIOLOGY, 2010).

E sabendo que regiões homólogas costumam compartilhar similaridade, as ferramentas de clusterização também são essenciais quando se deseja entender determinadas regiões da sequência que ainda não receberam uma classificação com base em suas relações evolutivas como a do Pfam (FINN, et al., 2016) ou SCOP (MURZIN, et al., 1995; ANDREEVA, et al., 2007).

2.4 Anotação de proteínas baseada em inferência de homologia

A inferência de homologia é uma das abordagens mais utilizadas para anotação de sequências em larga escala, já que a homologia é muitas vezes definida pela similaridade estrutural e/ou de sequência. Neste tipo de abordagem a ocorrência de alta similaridade entre duas sequências de proteínas é considerada um bom indicativo de ancestralidade (PEARSON, 2013; PUNTA; MISTRY, 2016).

A busca de similaridade de sequências para identificar sequências homólogas é uma das primeiras e mais informativas dentre as etapas na análise de sequências. Vários programas de busca de similaridade como BLAST (ALTSCHUL, 1990), PSI-BLAST (ALTSCHUL, 1990) e HMMER3 (SEAN; WEELHER, 2018) produzem alinhamentos com estatísticas precisas, atestando que sequências que compartilham alta similaridade também possuem estruturas similares. Assim, sequências que compartilham similaridade com scores significativos podem ser

inferidas como homólogas, ou seja, é possível dizer que elas compartilham um ancestral comum (PEARSON, 2013). Porém, sequências homólogas nem sempre compartilham alta similaridade de sequência, pois existem proteínas homólogas que compartilham alta similaridade estrutural, mas não apresentam alinhamentos com escores estatísticos significantes (LOEWENSTEIN, et al., 2009; PEARSON, 2013).

Hoje existem muitas fontes públicas que disponibilizam conjuntos de dados de famílias de proteínas que podem ser utilizados na inferência de homologia e anotação de sequências biológicas. Alguns exemplos deste tipo de fonte são os bancos de dados que classificam as proteínas de acordo com os seus domínios proteicos e arquiteturas conservadas, como Pfam (FINN, et al., 2016), SMART (LETUNIC; BORK, 2018), TIGRFAMS (HAFT, 2013) e PANTHER (HUAIYU, et al.,2013; HUAIYU, et al.,2015; HUAIYU, 2016).

2.5 Banco de Dados Pfam

O banco de dados Pfam (FINN, et al., 2016) é considerado um dos mais generalistas devido ao seu formato de classificação das regiões conservadas. Em sua última atualização (31.0) que ocorreu em março de 2017, o banco apresentava no total 16.172 famílias e 604 clãs, sendo que nesta data foram construídas 415 novas famílias, 11 novos clãs e 9 famílias foram eliminadas.

As famílias do Pfam são criadas com base nos alinhamentos múltiplos das sequências representantes, ou seja, as sequências as quais acredita-se serem homólogas são alinhadas com as demais e o resultado deste alinhamento é utilizado para treinar os modelos ocultos de Markov (HMM) que serão associados a cada uma das famílias criadas (FINN, et al., 2016; PUNTA; MISTRY, 2016).

O modelo oculto de Markov, do inglês *Hidden Markov Models*, é um modelo estatístico muito utilizado na Bioinformática devido a sua capacidade de extrair do alinhamento múltiplo de sequências atributos que podem caracterizar as famílias de proteínas, como por exemplo a probabilidade de ocorrerem inserções, deleções e substituições de aminoácidos em posições específicas na sequência (DURBIN, et. al., 1998; GHAHRAMANI, 2001; EDDY, 1998; EDDY; PEARSON, 2001; GUNASEKARAN, 2017; KROGH, et. al., 1994).

Hoje o Pfam utiliza o pacote de ferramentas HMMER3 (SEAN; WEELHER, 2018) para criar os HMM’s que representam as famílias de proteínas. As sequências

depositadas são classificadas em 6 grupos (FINN, et al., 2016):

I. Família: uma coleção de regiões de proteínas que apresentam relação entre si com base nos alinhamentos e HMM's;
II. Domínio: uma unidade estrutural da proteína;
III. Unidade Polipeptídica de Repetição: uma unidade estrutural curta que é instável quando isolada, porém estável quando várias cópias estão presentes na proteína;
IV. Motivos: uma unidade curta encontrada fora dos domínios globulares;
V. *Coiled-coils*: regiões que apresentam predominância de motivos do tipo *coiled-coil,* que são motivos em que geralmente 2 a 7 alfa-hélices se encontram superenroladas, sendo assim também conhecido como superhélice;
VI. Região Desordenada: regiões conservadas que são intrinsecamente desordenadas e não globulares.
VII. Clãs: As proteínas agrupadas em Famílias podem ainda ser distribuídas em Clãs se apresentarem uma possível relação evolutiva.

2.5.1 Clãs

No Pfam um Clã é definido como um grupo que contém duas ou mais famílias que provavelmente surgiram de uma origem evolutiva comum (FINN, et al., 2006; FINN, et al., 2016). Esta ancestralidade é inferida por meio de quatro evidências independentes:

I. Similaridade de sequência
II. Similaridade de estrutura tridimensional já resolvida experimentalmente
III. Similaridade Funcional
IV. Similaridade entre os HMMs

É importante ressaltar que de acordo com a literatura a presença de estruturas 3D relacionadas, scores significativos na comparação entre os HMM e ocorrência de *E-value* entre 0.1 e 0.001 são considerados bons indicadores primários de relações entre as famílias. Portanto, os clãs do Pfam podem ser

facilmente utilizados para relacionar sequências com função desconhecida a famílias de proteínas que já foram bem caracterizadas (FINN, et al., 2006; FINN, et al., 2016; PUNTA; MISTRY, 2016).

2.5.2 PfamScan

Existem dois programas do pacote HMMER3 (SEAN; WEELHER, 2018) que realizam a busca de Famílias do Pfam em um dado conjunto de sequências. Um deles, o *hmmsearch* faz a busca de um HMM contra um conjunto de sequências. O outro, chamado *hmmscan,* busca uma sequência contra um conjunto de HMM's (SEAN; WHEELER, 2018).

O *hmmscan* foi utilizado pelo Pfam (FINN, et al., 2016) para criar um script denominado *pfam_scan.pl*. Este script foi desenvolvido na linguagem Perl® e pode ser obtido no website do Pfam. Hoje de acordo com a literatura, uma das formas de classificar um conjunto de dados de acordo com as anotações do Pfam é utilizar o programa PfamScan (LI, et al., 2015; MISTRY; BATEMAN; FINN, 2007; PUNTA; MISTRY, 2016).

2.6 Banco de Dados *Gene Ontology* e Anotação Funcional

O *Gene Ontology* (GO) é um banco de dados que disponibiliza informações funcionais de produtos gênicos como genes, proteínas e complexos macromoleculares utilizando ontologias baseadas em evidências (ASHBURNER et. al., 2000; BALAKRISHNAN, et. al., 2013).

O Gene Ontology provê a descrição de genes e seus produtos gênicos de forma consistente em qualquer organismo, além de produzir uma plataforma adaptada para processar os dados a nível funcional. A estrutura do Gene Ontology é formada por 3 hierarquias organizadas como um gráfico acíclico direcionado (DAG): *Biological Process* (BP), *Molecular Function* (MF) e *Cellular Component* (CC). No DAG cada aresta representa um tipo de relação entre os termos GO que são semântica e topologicamente unidos pelas seguintes relações: '*is a*', '*part of*', '*regulates*' ('*positively regulates*' /' *negatively regulates'*), '*has part*' ou '*occurs in*' (MAZANDU; MULDER, 2013; MAZANDU, et al., 2015; MAZANDU; CHIMUSA; MULDER, 2017).

Entretanto, é importante ressaltar que as relações de hierarquia '*is a*' e '*part of*' formam a parte principal das ontologias. Assim, estas relações associativas poderiam indicar, por exemplo, que um componente celular é o local aonde ocorre um determinado processo biológico (BODENREIDER; AUBRY; BURGUN, 2005). Assim, uma anotação do GO descreve as associações entre uma classe da ontologia e um produto gênico. (ASHBURNER et. al., 2000; BALAKRISHNAN, et. al., 2013).

Atualmente o Consórcio *Gene Ontology* (GOC) adota duas classes de anotação: as que são manualmente revisadas por biocuradores experientes, e as que são criadas automaticamente por métodos computacionais (GOC; 2015).

Outra característica do banco de dados GO importante para a compreensão deste trabalho é a definição dos códigos de evidência (do inglês *Evidence Codes, EC*). Sabendo que uma anotação funcional do GO representa um termo GO associado a uma referência bibliográfica específica que descreve a análise que deu origem a associação entre um termo GO e um produto gênico, todas as anotações incluem um código de evidência. Os códigos de evidência indicam a literatura na qual o termo GO foi embasado (GOC; 2004; GOC; 2018). Dentre os códigos de evidência (FIGURA 2) apenas o código IEA (*Electronic Annotation*) não é nomeado por um curador humano.

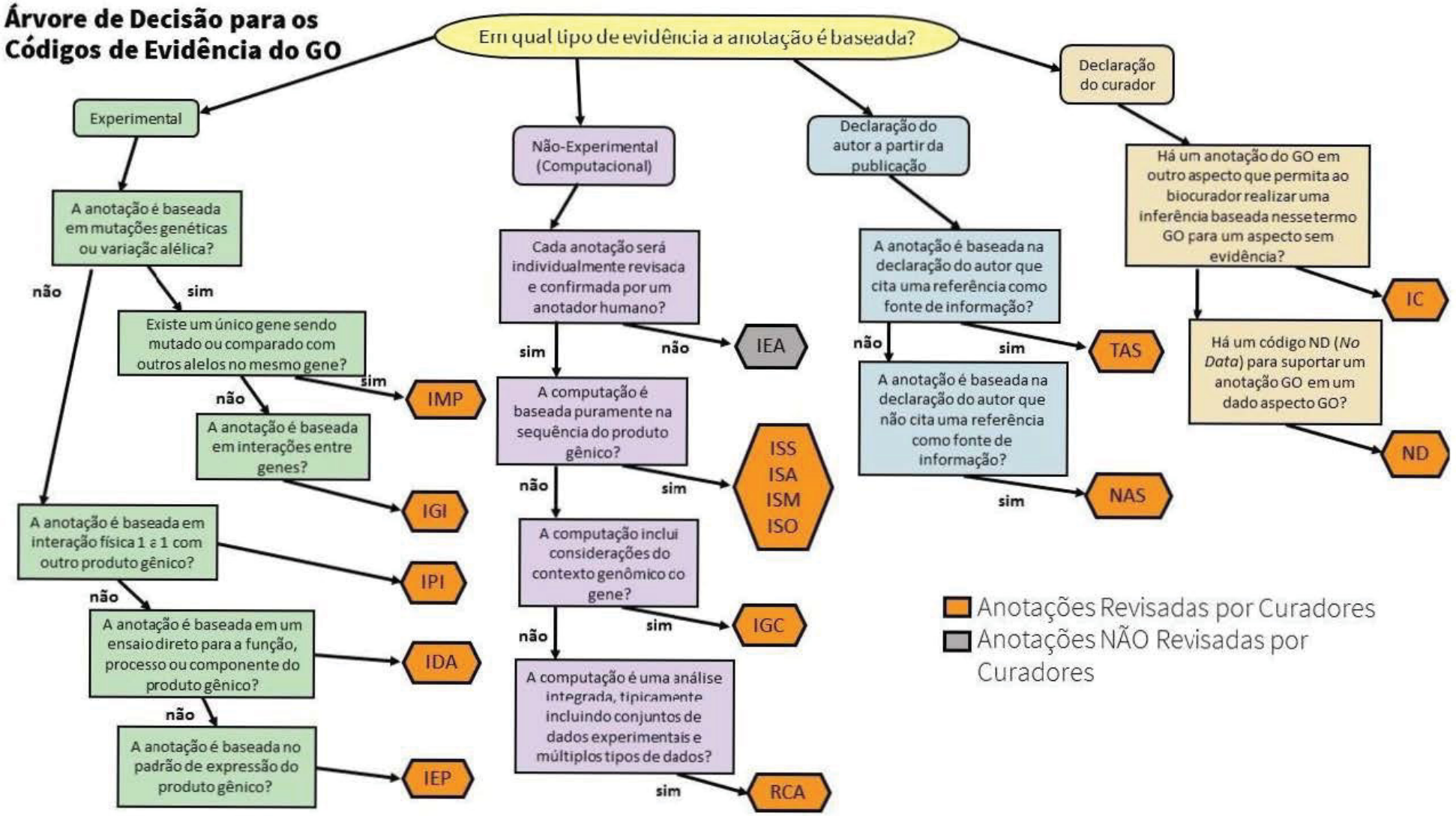
Existem 6 grupos de códigos de evidência no GO:

I. Experimentais: indica que a referência bibliográfica citada na anotação é resultado de uma caracterização física de um gene ou produto gênico;
II. Sequênciamento de Alta Vazão: podem ser utilizados para anotação baseada em metodologias como o sequênciamento de alta vazão;
III. Análise Computacional: indicam que a anotação é baseada em resultados de análises realizadas *in silico*;
IV. Declaração do autor: a anotação foi realizada com base em uma declaração do autor(es) da referência bibliográfica do termo GO;
V. Declaração do curador: indica que a anotação foi produzida com base no discernimento do curador do GO;
VI. Anotação Eletrônica: é atribuído por métodos automatizados, sem a intervenção de um curador humano.

Assim, os códigos de evidência conseguem diferenciar anotações baseadas

em experimentos realizados em laboratório dos realizados *in silico* por meio de predições (GOC, 2018; ROGERS; BEM-HUR, 2009).

FIGURA 2 - ARVORE DE DECISÃO PARA OS CÓDIGOS DE EVIDÊNCIA DO GENE ONTOLOGY



FONTE: A Autora (2018). Adaptado de Gene Ontology Consortium (2018).

3 JUSTIFICATIVA

A imensa quantidade de dados biológicos, como sequências de aminoácidos, anotações funcionais depositadas, e artigos científicos publicados nos últimos anos, além do índice crescente de bases de dados como o SwissProt e o *Non-redundant Data Sequences of Proteins* (NR), tornou humanamente impossível a extração manual de todas as relações biológicas presentes tanto na literatura, quanto nos bancos de dados biológicos. Assim, para solucionar este problema e conseguir associar os termos de importância biológica, diversas técnicas de mineração de dados têm sido desenvolvidas (CHAWLA; SHARMA, 2016; LI; JAROSZEWSKI; GODZIK, 2002), sendo que dentre elas a clusterização de bancos de sequências biológicas têm se destacado por ser capaz de reduzir o tamanho, acelerar a busca contra os bancos, além de relevar informações biológicas relevantes sobre determinados conjuntos de dados.

Entretanto, em 2018, Chen e colaboradores (CHEN, et. al., 2018) levantaram questionamentos acerca do estado da arte na clusterização de bancos de dados de sequências biológicas e chegaram à conclusão de que os maiores *clusters* de proteínas gerados após a clusterização são casos que precisam ser cuidadosamente estudados, sendo muitas vezes necessário o reprocessamento dos resultados. Porém, a melhor forma de validar e interpretar biologicamente e verificar a coesão *intracluster* nos maiores clusters produzidos ainda permanece sem resposta na comunidade, pois na literatura ainda não há recomendações práticas de como tratar os resultados da clusterização de um banco de dados de sequências biológicas.

Diante disso, este trabalho propôs a criação de um pipeline para analisar “big clusters”, ou seja, selecionar grandes grupos de proteínas e aplicar técnicas de mineração de dados como a mineração de textos e a clusterização para extrair e descobrir informações relevantes sobre estes dados que permitam sua interpretação e integração, além de produzir recomendações práticas aos usuários de ferramentas de clusterização.

4 OBJETIVOS

4.1 Objetivo Geral

Analisar grandes grupos de proteínas utilizando técnicas de mineração de dados e criar um pipeline reprodutível para reprocessamento e interpretação de resultados obtidos na clusterização de bancos de dados de sequências biológicas.

4.2 Objetivos Específicos

Utilizando os 21 maiores *clusters* gerados na clusterização do banco de dados NR, os objetivos específicos deste trabalho foram:

I. Validar, integrar e interpretar estes *clusters*

II. Tornar os dados úteis para os usuários finais da informação

III. Descobrir padrões, tendências e relacionamentos interessantes nestes dados

IV. Descobrir quais famílias de proteínas estão presentes nos conjuntos de dados

V. Classificar os *clusters* de acordo com a (s) função (ões) biológicas das proteínas que os compõem

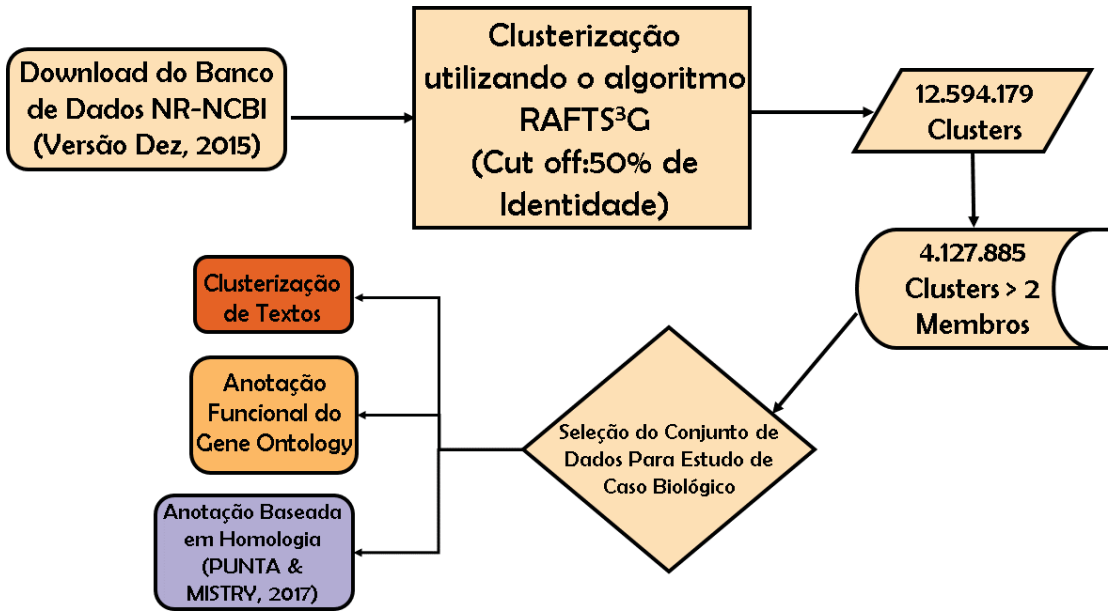
VI. Descobrir se existem relações de homologia entre as sequências que compõem o conjunto de dados

VII. Intitular os *clusters* de acordo com seu conteúdo e sentido biológico

5 MATERIAIS E MÉTODOS

Este trabalho foi desenvolvido de acordo com o fluxograma apresentado nas FIGURAS 3, 4, 5, 6, 7 e 8. Cada legenda descreve as etapas realizadas na metodologia e quais materiais foram utilizados.

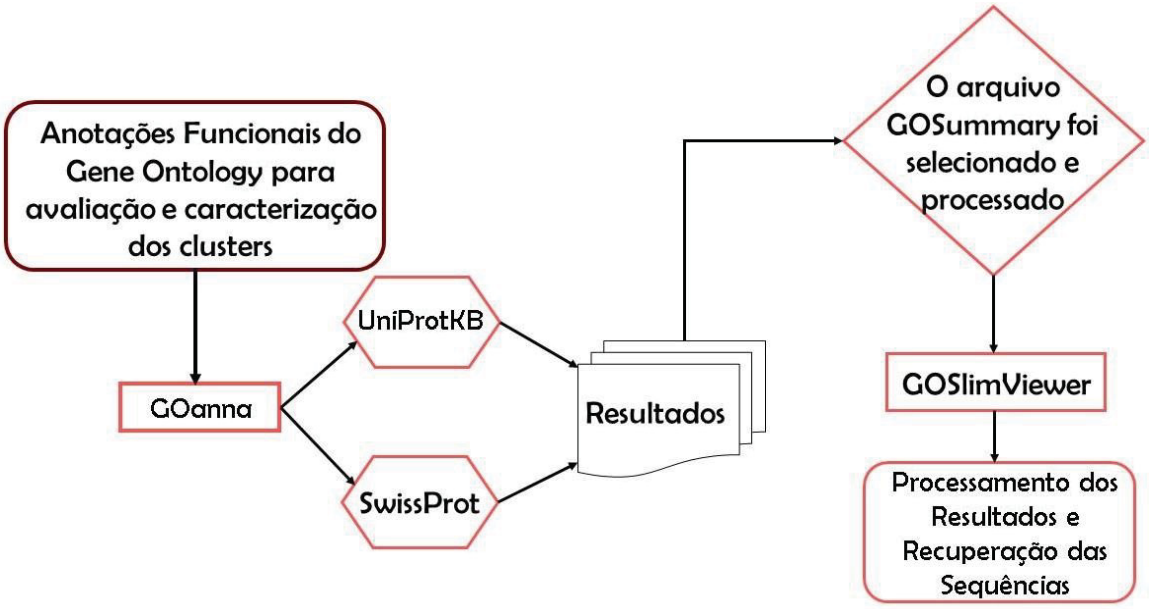
FIGURA 3 - PARTE A DO FLUXOGRAMA REPRESENTANDO OS MATERIAIS E A METODOLOGIA UTILIZADOS NO TRABALHO



FONTE: A Autora (2018).

LEGENDA: Fluxograma da primeira etapa do trabalho. Este trabalho foi desenvolvido sob duas perspectivas: Estudo de caso biológico e criação do pipeline. Os conjuntos de dados escolhidos para o estudo de caso biológico são os 21 maiores clusters obtidos na clusterização do banco de dados NR com valor de corte de 50% de identidade.

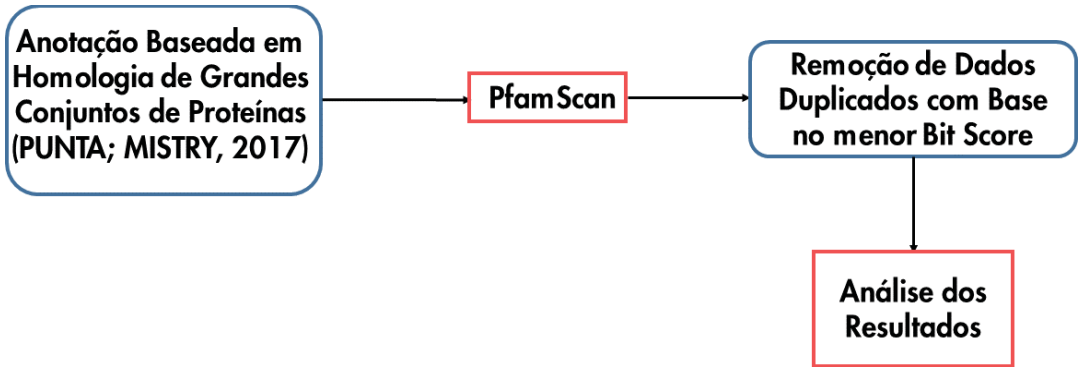
FIGURA 4 - PARTE B DO FLUXOGRAMA REPRESENTANDO OS MATERIAIS E A METODOLOGIA UTILIZADOS NO TRABALHO



FONTE: A autora (2018).

LEGENDA: Na segunda etapa do trabalho, com o intuito de realizar o estudo de caso biológico e gerar o pipeline foram realizadas análises funcionais baseadas nas anotações funcionais do Gene Ontology.

FIGURA 5 - PARTE C DO FLUXOGRAMA REPRESENTANDO OS MATERIAIS E A METODOLOGIA UTILIZADOS NO TRABALHO



FONTE: A autora (2018).

LEGENDA: Na terceira etapa do trabalho, com o intuito de realizar o estudo de caso biológico e gerar o pipeline foram realizadas análises com base na metodologia proposta nos trabalhos de PUNTA & MISTRY (PUNTA; MISTRY, 2016).

5.1 Download do Banco de Dados NCBI *Non-redundant Data Sequences of Proteins* (NR)

Inicialmente, nós obtivemos todas as sequências de aminoácidos depositadas no banco de dados *Non-redundant Data Sequences of Proteins* (NR) do NCBI (XU, 2012). O arquivo escolhido para download estava no formato FASTA e continha 78.002.046 sequências. A versão utilizada foi a de dezembro de 2015.

5.2 Clusterização do Banco de Dados *Non-redundant Data Sequences of Proteins* (NR) utilizando o algoritmo RAFTS^{3}G

Após a obtenção dos dados foi realizada a clusterização das sequências de aminoácidos utilizando o algoritmo RAFTS^{3}G (*Rapid Alignment Free Tool for Sequences Similarity Search*) (NICHIO, et al., 2018; NICHIO, 2016; VIALLE, 2013).

O valor de self-score escolhido para clusterização do banco foi de 50% pois de acordo com a literatura (PEARSON,2013) é possível inferir homologia entre sequências de aminoácidos que compartilhem de no mínimo 50% de identidade.

5.2.1 RAFTS^{3}G

RAFTS^{3}G é um algoritmo aplica uma abordagem livre de alinhamento para clusterizar as sequências biológicas com base na ferramenta RAFTS3 (VIALLE, et al., 2016). Assim, dado um arquivo FASTA como entrada no RAFTS^{3}G, a primeira etapa, que é realizada pelo RAFTS3, é composta por dois passos principais de formatação:

I. Indexação das sequências por meio de uma função de *hash* (ALTSCHUL, et al., 1990; BUCHFINK; XIE; HUSON, 2015)*;*

II. Atribuição de uma matriz de coocorrência de aminoácidos (BCOM) a cada sequência com o objetivo de representar as sequências com poucos *bytes* de memória.

Assim, no primeiro passo um conjunto de *k-mers* (KIM, et al., 2017; PAN, et al., 2017) é aleatoriamente selecionado para cada uma das sequências do arquivo

FASTA e em seguida, submetido a uma função de *hash*. A tabela de *hash* e a matriz BCOM são então salvas na mestra estrutura. Com esta estrutura temos então o banco formatado e, portanto, é possível realizar a etapa de busca por similaridade.

A etapa de busca por similaridade é realizada em duas partes principais: filtragem e comparação. Durante a filtragem, é realizada a seleção por meio da tabela de *hash* apenas das sequências que contém *k-mers* em comum com a sequência *query*. Estas sequências que compartilham um mesmo índice de *k-mers* são as sequências candidatas. A fase de comparação e cálculo de similaridade é realizada entre as sequências candidatas por meio da soma binária de suas respectivas matrizes BCOM. Cada sequência candidata selecionada também recebe uma medida de dissimilaridade. Os melhores resultados passam então por um alinhamento global (SMITH; WATERMAN, 1981).

Após esta etapa de formatação do banco pela ferramenta RAFTS3, o RAFTS^{3}G analisa cada sequência como um *cluster* em potencial e realiza a busca de 50 sequências aleatórias contra o banco e avalia a identidade entre as sequências por meio de resultado do alinhamento com realizado com base no *self-score* dado pelo usuário. Em seguida os clusters são criados com base nestes resultados.

5.3 Seleção do Conjunto de Dados para Caso de Estudo Biológico

Após a clusterização do banco *Non-redundant Data Sequences of Proteins (NR)* foram obtidos 12.594.179 *clusters*, sendo 4.127.885 com mais de 2 membros e 8.466.294 únicos. Analisando a distribuição das sequências de aminoácidos nos *clusters* observou-se que apenas 21 continham mais de 100.000 sequências e somente um deles mais de 500.000 sequências. Assim sendo, os 21 maiores *clusters* foram selecionados para realizar um estudo de caso biológico.

5.4 Análises Funcionais

Após a escolha do conjunto de dados que seria utilizado para o estudo de caso biológico, duas análises foram realizadas: uma baseada nas anotações do *Gene Ontology* e outra baseada em inferência de homologia em grandes grupos de

proteínas.

5.4.1 Análise Baseada na Anotação Funcional do Banco *Gene Ontology*

5.4.1.2 Obtenção das Anotações do *Gene Ontology*

Para caracterizar o conjunto de dados com base nas anotações do *Gene Ontology* (ASHBURNER et. al., 2000) a primeira ferramenta escolhida foi a GOanna (McCARTHY, et al., 2006). Esta ferramenta transfere as anotações do Gene Ontology (ASHBURNER et. al., 2000) com base em homologia de sequências por meio de um padrão do BLASTp (ALTSCHUL, et al., 1990; CAMACHO, et al., 2009) contra bancos de dados que contém sequências anotadas no Gene Ontology (ASHBURNER et. al., 2000).

Nesta análise foram utilizados os seguintes parâmetros na ferramenta GOanna (McCARTHY, et al., 2006):

I. Bancos de Dados Utilizados:
    a. UniProt (The UniProt Consortium, 2017)
    b. SwissProt (The UniProt Consortium, 2017)
II. *Expect*: $10^{-e\ 50}$
III. *Matrix*: BLOSUM62
IV. *Wordsize*: 6
V. *GapCosts*: *Existence 11: Extension: 1*
VI. Número de Sequências Alvo: 5
VII. Filtro de Baixa Complexidade: Sim
VIII. Porcentagem de Identidade: 60 %
IX. Filtro de Cobertura da Sequência Query: 60%
X. Tipos de Códigos de Evidência:
    a. Códigos de Evidências Experimentais: EXP, IDA, IPI, IMP, IGI, IEP, HTP, HGI, HDA, HEP)
    b. Códigos Homologados por Curadoria (IC, ND)
    c. Códigos de evidência gerados automaticamente foram excluídos dessa análise

5.4.1.3 Processamento e Sumarização dos Resultados obtidos com a ferramenta GOanna

Foram obtidos os seguintes arquivos de saída na ferramenta GOanna (McCARTHY, et al., 2006):

I. **Align.html**: Contém os resultados do alinhamento realizado pelo BLASTp (ALTSCHUL, et al., 1990)
II. **Arquivo do Excel**: contém os arquivos de entrada, as sequências as quais foram realizados os alinhamentos, e as anotações do *Gene Ontology* (ASHBURNER et. al., 2000) para os hits que passaram pelos filtros de *E-value*, Porcentagem de identidade e Cobertura da Sequência de Entrada.
III. **sliminput.txt**: resume os resultados com base no conjunto de dados *Gene Ontology Slim* (GENE ONTOLOGY CONSORTIUM, 2008; BISWAS, et al., 2002).
IV. **matches.txt**: este arquivo é gerado para entradas com ID desconhecido.

O arquivo GO*Summary* foi selecionado e utilizado como entrada na ferramenta GO*SlimViewer* (McCARTHY, et al., 2006). O GO*SlimViewer* (McCARTHY, et al., 2006) é uma ferramenta linha de comando que fornece um resumo de alto nível das anotações do *Gene Ontology* (ASHBURNER et. al., 2000) de um conjunto de dados. A ferramenta GOSlimViewer (McCARTHY, et al., 2006) foi utilizada com os seguintes parâmetros:

I. Conjunto de Dados GO Slim: Generic;
II. Biblioteca: Perl5;

A chamada do script *goslimviewer_standalone.pl* foi realizada da seguinte forma:

perl goslimviewer_standalone.pl -i input_text_file -s slim_dataset(generic,metagenomics,panther,goa,pir,plant,tigr,yeast) [-o output_file_prefix]

5.4.2 Análise *Baseada* em Inferência de Homologia em Grandes Grupos de Proteínas

Para realizar a anotação baseada em Inferência de Homologia em Grandes Grupos de Proteínas foi adotada a metodologia sugerida nos trabalhos de Punta & Mistry (PUNTA; MISTRY, 2016).

5.4.2.1 PfamScan

Então o arquivo FASTA do *Cluster* n° 1 foi utilizado como entrada na ferramenta *PfamScan* (LI, et al., 2015; MISTRY; BATEMAN; FINN, 2007). O *PfamScan* (LI, et al., 2015; MISTRY; BATEMAN; FINN, 2007) faz a busca de uma sequência contra um conjunto de modelos ocultos de Markov (HMMs). Nesta análise foram utilizados o conjunto de perfis HMM Pfam-A, devido ao fato de que este é o conjunto que foi manualmente curado por especialistas.

Assim, nesta etapa foram utilizados os parâmetros padrão do algoritmo de acordo com os seguintes passos:

I. Obtenção do algoritmo *pfam_scan.pl* no FTP do Pfam (LI, et al., 2015; MISTRY; BATEMAN; FINN, 2007);
II. *Download* dos HMMs do conjunto Pfam-A (*Pfam-A.hmm*) e do arquivo associado (*Pfam-A.dat*);
III. Chamada do script *pfam_scan.pl* por linha de comando da seguinte forma:

pfam_scan.pl –fasta <FASTA format sequence filename>-dir <location of Pfam profile-HMM files>><output filename>

O arquivo de saída do *PfamScan* (FIGURA 6) contém informações como as

divisões das famílias com as quais houve alinhamento para cada uma das sequências, as coordenadas do alinhamento e do envelope, número de acesso, nome e tipo do HMM, a região do perfil HMM que alinhou com a(s) sequência(s) em determinadas regiões, o tamanho do modelo HMM, *bit-score* e *E-value* do alinhamento, significância do hit (0 ou 1) e em qual Clã a sequência foi classificada.

FIGURA 6 - ARQUIVO DE SAÍDA DO PFAMSCAN

```
# <seq id> <alignment start> <alignment end> <envelope start> <envelope end> <hmm acc> <hmm name> <type> <hmm start> <hmm end>
hmm length> <bit score> <E-value> <significance> <clan>

gi|446081699|ref|WP_000159554.1|          29    187     28    188 PF00005.26  ABC_tran          Domain     2   136   137
102.0   3.3e-29   1 CL0023
gi|446081699|ref|WP_000159554.1|         239    301    239    302 PF08352.11  oligo_HPY         Family     1    63    65
42.2   8.2e-11   1 No_clan
gi|488941443|ref|WP_002852518.1|          73    251     61    303 PF00528.21  BPD_transp_1      Family     5   178   185
62.9     3e-17   1 CL0404
gi|447202750|ref|WP_001280006.1|          19    182     19    182 PF01765.18  RRF               Domain     1   162   162
206.3   2.4e-61   1 No_clan
gi|446740104|ref|WP_000817360.1|          22    105     21    106 PF04341.11  DUF485            Family     2    88    89
84.1   4.6e-24   1 No_clan
gi|489815775|ref|WP_003719603.1|           8     88      8     88 PF00312.21  Ribosomal_S15     Domain     1    81    81
124.7   1.2e-36   1 CL0600
gi|24660781|ref|NP_648201.1|              10    176      4    176 PF00025.20  Arf               Domain     7   175   175
222.2   2.8e-66   1 CL0023
gi|490291073|ref|WP_004186678.1|           4    240      4    241 PF00483.22  NTP_transferase   Family     1   246   248
242.2   5.7e-72   1 CL0110
gi|488137897|ref|WP_002209105.1|           1    238      1    239 PF00370.20  FGGY_N            Domain     2   244   245
124.3   5.3e-36   1 CL0108
gi|488137897|ref|WP_002209105.1|         248    430    248    436 PF02782.15  FGGY_C            Domain     1   192   198
91.8   4.3e-26   1 CL0108
gi|497571435|ref|WP_009885619.1|          34     96     34     96 PF00137.20  ATP-synt_C        Family     1    60    60
50.2   2.3e-13   1 No_clan
gi|446046984|ref|WP_000124839.1|           3    109      3    109 PF00453.17  Ribosomal_L20     Family     1   107   107
165.4   3.2e-49   1 No_clan
gi|490007860|ref|WP_003910692.1|           8    164      8    165 PF00823.18  PPE               Family     1   157   158
195.0   7.6e-58   1 CL0352
gi|446969437|ref|WP_001046693.1|           8     62      7     62 PF14056.5   DUF4250           Family     2    55    55
74.2   5.2e-21   1 No_clan
gi|489346062|ref|WP_003253197.1|           6     71      6     71 PF02874.22  ATP-synt_ab_N     Domain     1    69    69
76.1     2e-21   1 CL0275
gi|489346062|ref|WP_003253197.1|         128    340    128    340 PF00006.24  ATP-synt_ab       Domain     1   213   213
214.9   9.2e-64   1 CL0023
gi|445975466|ref|WP_000053321.1|          33     77     33     78 PF04226.12  Transgly_assoc    Family     1    48    49
40.4   2.4e-10   1 No_clan
gi|499956927|ref|WP_011637661.1|           5    144      4    144 PF00075.23  RNase_H           Domain     2   143   143
164.0   2.2e-48   1 CL0219
gi|489184935|ref|WP_003094362.1|           5     83      3     84 PF00381.18  PTS-HPr           Domain     4    81    82
```

FONTE: A Autora (2018).

LEGENDA: Arquivo de saída do algoritmo *pfam_scan.pl*. As colunas 2 e 3 representam as coordenadas do alinhamento, ou seja, as regiões nas quais o HMMER (SEAN; WEELHER, 2018) conseguiu realizar um alinhamento com o perfil HMM, enquanto as colunas 4 e 5 representam as coordenadas do envelope, ou seja, a extensão do alinhamento com um homólogo que o HMMER (SEAN; WEELHER, 2018) conseguiu identificar mesmo sem alinhar todos os resíduos da região. As colunas 6 a 8 apresentam o número de acesso, nome e tipo do HMM. As colunas 9 a 11 mostram a região do perfil HMM que alinhou com a(s) sequência(s) em determinadas regiões e o tamanho do modelo HMM. As colunas 12 e 13 correspondem respectivamente, ao *bit-score* e ao *E-value* do alinhamento. A coluna 14 contém a significância do hit (0 ou 1). A coluna 15 apresenta a qual Clã do Pfam a sequência pertence.

Esta etapa forneceu uma grande quantidade de dados não organizados e que necessitavam de processamento para utilização e extração de informações. Assim, afim de facilitar a manipulação e interpretação destes dados, eles foram armazenados em um banco de dados utilizando a biblioteca SQLITE3® (HIPP, et al., 2015) do Python3® (ROSSUM; DRAKE, 2009) e a ferramenta MATLAB® como IDE.

5.5 Mineração de Texto

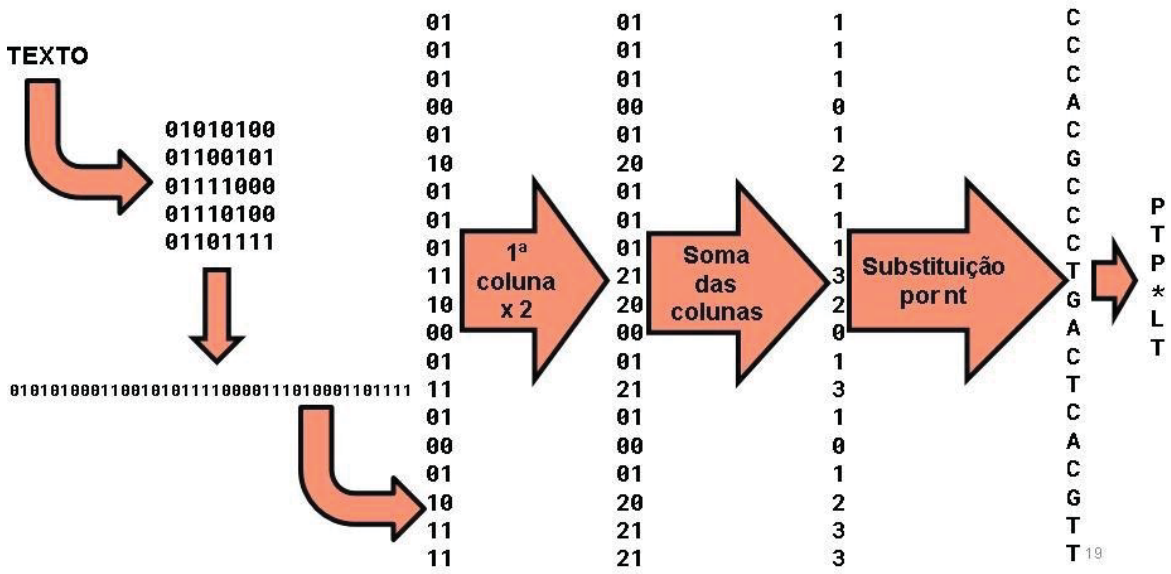
Para nomear os clusters com base em seu conteúdo, realizamos também a aplicação da metodologia de clusterização e mineração de textos desenvolvida pelo nosso grupo de pesquisa no arquivo FASTA correspondente ao *cluster* n° 1 do bando de dados NR.

5.5.1 Pré-processamento dos Dados para Aplicação de Mineração de Texto e Denominação dos *Clusters*

Inicialmente, para preparar o conjunto de dados para aplicar a mineração de texto, apenas o cabeçalho do arquivo FASTA foi extraído e codificado para o formato de nucleotídeos (timina, adenina, citosina e guanina). Com o cabeçalho representado no formato de nucleotídeos, foi realizada a transformação para aminoácidos utilizando a função *dna2list* (APÊNDICE 2). Portanto, nesta etapa (FIGURAS 7 e 8) obteve-se um novo arquivo FASTA contendo a representação do Cabeçalho do FASTA do *cluster* n° 1 em formato de aminoácidos. Todos estes passos fazem parte da função *getheadersfeats* (APÊNDICE 3).

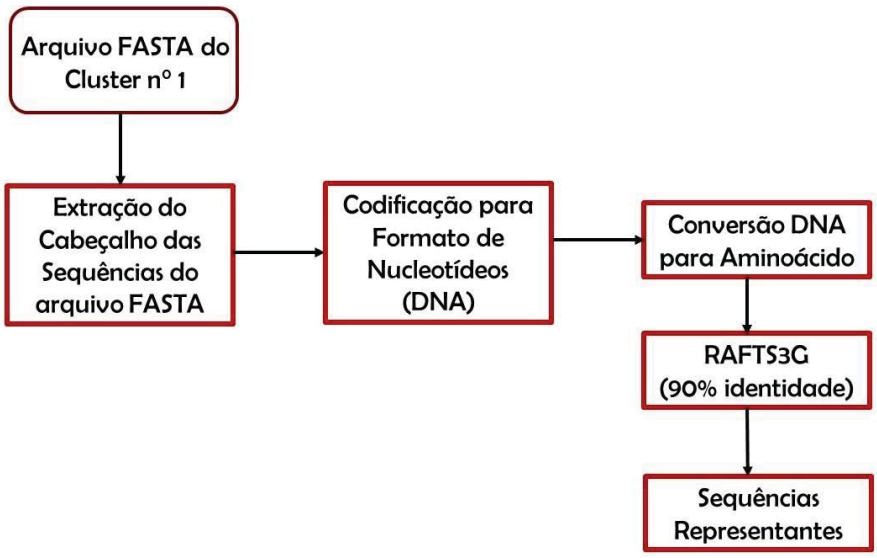
Então foi realizada a clusterização utilizando o algoritmo RAFTS^{3}G (NICHIO, et al., 2018; NICHIO, 2016) com *self-score* de 90% de identidade (FIGURA 8). Para diminuir o tempo de busca contra o conjunto de dados clusterizado e facilitar a interpretação dos dados a partir do nome padrão dados aos *clusters*, somente as sequências representantes de cada um dos *clusters* foram selecionadas.

FIGURA 7 - PARTE D DO FLUXOGRAMA REPRESENTANDO OS MATERIAIS E A METODOLOGIA UTILIZADOS NO TRABALHO



FONTE: A Autora (2018).
LEGENDA: Transformação dos textos em aminoácidos. Inicialmente é feita a transformação do texto em vetores de 8 *bits* e logo após para vetores de *bytes*. Então é realizada a concatenação das linhas da matriz em apenas um vetor. Este vetor é então transposto para o formato de duas colunas e em seguida a multiplicação da primeira coluna pela segunda, e então a soma dos valores da primeira com os valores da segunda coluna. Temos neste ponto quatro possibilidades de resultado: 0, 1, 2 e 3. Cada um destes números representa um nucleotídeo (A, T, C e G): 0 = adenina; 1 = citosina; 2 = guanina; 3 = timina. Com isto teremos uma sequência de nucleotídeos que serão então traduzidos para o formato de aminoácidos por meio do conjunto de funções da Toolbox de Bioinformática do MATLAB®.

FIGURA 8 – PARTE E DO FLUXOGRAMA REPRESENTANDO O PRÉ-PROCESSAMENTO PARA APLICAÇÃO DA MINERAÇÂO DE TEXTO



FONTE: A autora (2018).
LEGENDA: Na penúltima etapa do trabalho, com o intuito de preparar os dados para aplicar a

mineração de texto foi realizado o pré-processamento do arquivo FASTA.

5.5.2 Denominação dos *Clusters*

Para denominar os *clusters* de acordo com o conteúdo dos Cabeçalhos dos arquivos FASTA foi produzido um script (APÊNDICE 4) em que somente os Cabeçalhos de todas as sequências dos *clusters* são vetorizados com base em *spaced words* (BODEN, et al., 2013; LEIMEISTER, et al., 2014) e projetados em uma base quasiortonormal de 729 pontos por meio da função *text2mat* (APENDICE 5). Após a obtenção dos vetores projetados é calculada a média da menor distância entre os centros dos vetores por meio da função *distminWC* (APENDICE 6). O vetor cujo índice na estrutura dos clusters (arquivo *contig2.mat*) seja igual a média obtida será escolhido como representante para dar o nome padrão aos *clusters*.

As funções e demais dependências deste script podem ser encontrados em: https://github.com/grazLet/Give-a-cluster-name.

6 RESULTADOS E DISCUSSÃO

Esta sessão foi dividida em clusterização do banco de dados *Non-redundant Data Sequences of Proteins* (NR), Análise de Homologia, Caracterização Funcional e Mineração de Texto. Os aspectos relevantes de cada subtópico serão discutidos dentro das sessões. Também é importante ressaltar que apenas os *clusters* com no mínimo 2 sequências foram incluídos nas análises realizadas neste trabalho.

6.1 Clusterização do Banco de Dados *Non-redundant Data Sequences of Proteins (NR),* redução do tamanho e distribuição dos clusters

A versão do banco de dados *Non-redundant Data Sequences of Proteins* em dezembro de 2015 continha 78.002.046 sequências dos bancos GenBank (BENSON, et al., 1999; BENSON, et al., 2012), UniProt (The UniProt Consortium, 2017), SwissProt (The UniProt Consortium, 2017), PIR (BARKER, et al., 2000), *Protein Research Foundation* (PRF) do Japão, e *Protein Data Bank* (BERMAN, et al., 2000). O arquivo FASTA de entrada tinha 46.7 GB de tamanho.

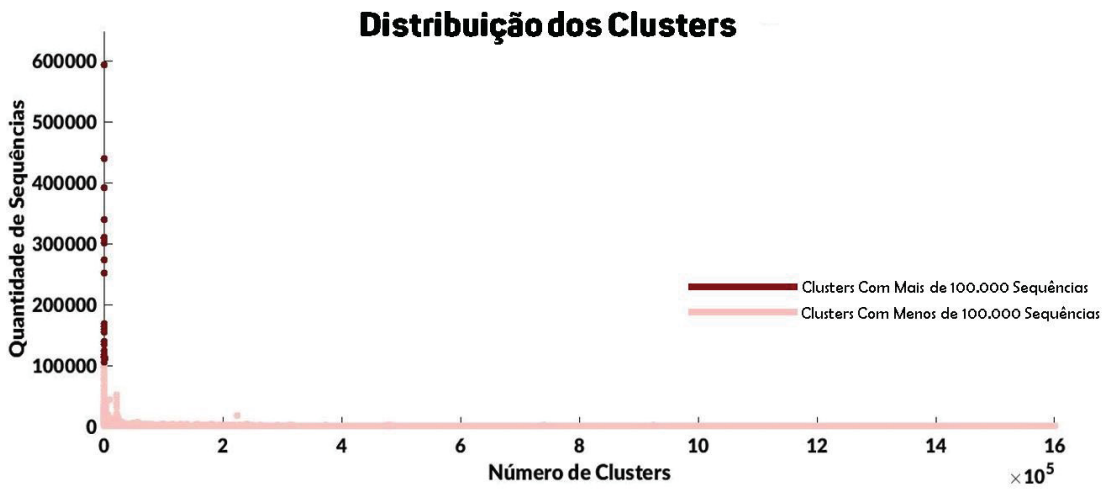
A clusterização do banco de dados *Non-redundant Data Sequences of Proteins* (NR) utilizando o algoritmo RAFTS^{3}G com valor de corte de 50% de identidade gerou 12.594.179 *clusters*, sendo que destes 4.127.885 continham mais de 2 sequências. O agrupamento levou cerca de 6 meses para ser concluído, tendo a estrutura de dados de todos os clusters gerados o tamanho de 447 MB o que demonstra a redução no tamanho do banco após a clusterização.

Analisando a distribuição das 78.002.046 sequências nos 4.127.885 clusters com mais de 2 sequências (FIGURA 8) observou-se que:

I. A média de sequências por cluster é 2

II. Apenas 21 clusters apresentaram mais de 100.000 sequências.

FIGURA 9 – DISTRIBUIÇÃO DOS CLUSTERS COM MAIS DE DUAS SEQUÊNCIAS



FONTE: A autora (2018).

Conforme será apresentado nas seções abaixo, para realizar um estudo de caso biológico selecionamos todos os clusters que continham mais de 100.000 sequências.

6.2 Anotação Baseada em Inferência de Homologia em Grandes Grupos de Proteínas

6.2.1 Resultados *PfamScan*

Sabendo que as anotações do banco de dados PFAM (FINN, et al., 2006; FINN, et al., 2016) podem ser muito úteis para interpretar um conjunto de dados de sequências biológicas, a etapa de anotação baseada em inferência de homologia seguiu os protocolos sugeridos nos trabalhos de Punta & Mistry (PUNTA; MISTRY, 2016). Na primeira etapa da análise a ferramenta *PfamScan* (LI, et al., 2015; MISTRY; BATEMAN; FINN, 2007) foi aplicada no arquivo fasta correspondente ao *cluster* n° 1.

Foi realizada a contagem da diversidade de Famílias, Domínios, Unidades Polipeptídicas de Repetição, *Coiled-coils* e Regiões Desordenadas de acordo com o banco de dados Pfam, conforme apresentado na TABELA 1. Nos resultados obtidos nenhuma sequência foi atribuída a classificação Motivo. É importante ressaltar que foi realizado um pré-processamento destes resultados para remover dados

duplicados de acordo com o menor valor de *bit-score* (APÊNDICE 1).

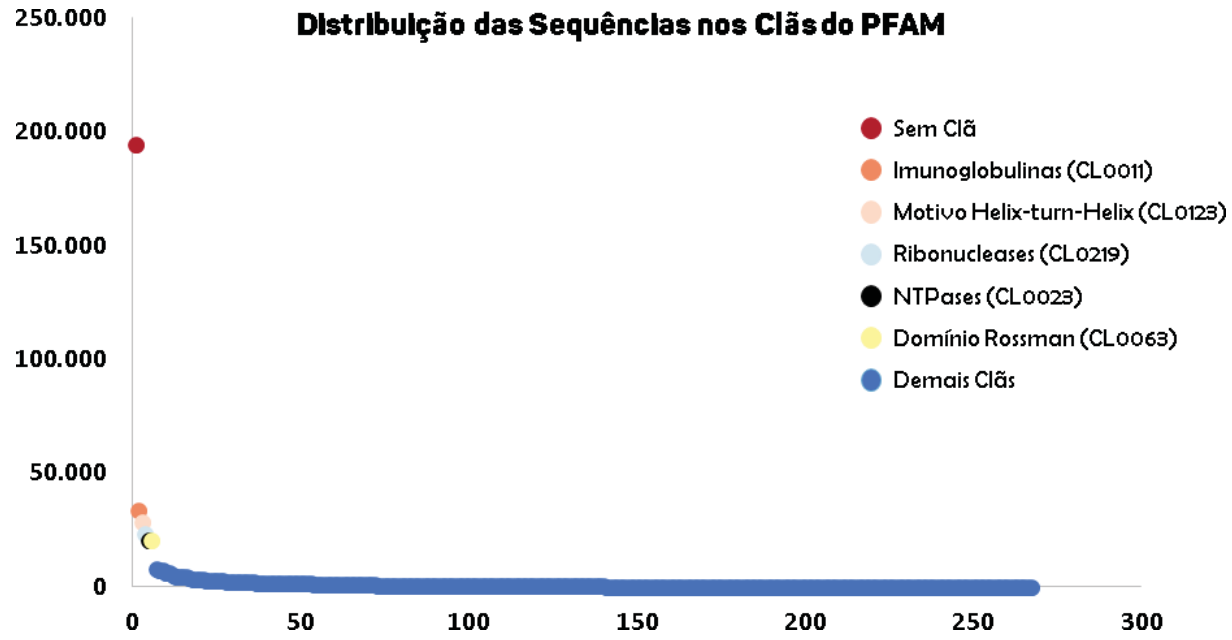
TABELA 1 - CLASSIFICAÇÕES DO PFAM NO *CLUSTER* 1

Famílias	Domínios	Unidades Polipeptídicas de Repetição	*Coiled-coils*	Regiões Desordenadas
1685	1377	55	8	1

6.2.1.1 Clãs

No caso de estudo biológico utilizando o arquivo FASTA do *cluster* nº 1, foi observado que 193.913 sequências não foram atribuídas a nenhum clã e 318.567 sequências foram atribuídas a 267 clãs diferentes (APÊNDICE 7), sendo que destes 5 se destacam por estar muito acima da média de distribuição de sequências por clã (1198 sequências) com mais de 10.000 sequências, conforme apresentado na FIGURA 10.

FIGURA 10 - DISTRIBUIÇÃO DAS SEQUÊNCIAS DO CLUSTER NÚMERO 1 NA CLASSIFICAÇÃO CLÃ DO PFAM



FONTE: A Autora (2018).

O maior clã encontrado no *cluster* nº 1, denominado Superfamília das Imunoglobulinas (do inglês *Immunoglobulin Superfamily)* contém 31 Famílias e 263.583 domínios proteicos. A superfamília das imunoglobulinas (IgSF) é uma grande família de proteínas de superfície caracterizadas pela presença de um domínio variável de cerca de 70 a 110 aminoácidos, incluindo uma grande variedade de receptores e funções (BUCK, 1992; HARPAZ; CHOTIA, 1994; McEVER; LUSCINSKAS, 2018; WILLIAMS; BARCLAY, 1988).

O clã CL0123, intitulado Motivo *Helix-turn-helix* (*Helix-turn-Helix clan* em inglês) contém 340 Famílias e 2215350 domínios. Este clã contém diversos domínios de ligação ao DNA que possuem o motivo *helix-turn-helix.* Este motivo é encontrado em todas as proteínas de ligação ao DNA que regulam a expressão genica (BRENNAN; MATTHEWS, 1989; PABO; SAUER, 1984; SAUER, et al., 1982), além de estar presente em todos os Domínios da Vida (ARAVIND, et al., 2005). Outra característica importante do domínio *helix-turn-helix* está no fato de que ele está presente em quase todos os fatores de transcrição bacterianos, bem como em um quarto dos fatores de transcrição humanos (ALQURASHI; TANG; XIA, 2015).

O terceiro maior clã encontrado no *cluster* nº 1 é o clã da Superfamília das Ribonucleases, também conhecidas como RNAses. Este grupo é composto por diversas nucleases que compartilham similaridade estrutural com a RNAse H e contém 61 Famílias e 316880 Domínios. As ribonucleases são enzimas presentes em humanos, eucariotos e procariotos responsáveis entre outras funções pela clivagem do RNA do duplex DNA-RNA durante os processos de replicação e reparo do DNA, exercendo papel fundamental em eucariotos superiores (HICE; CRAIGIE; DAVIES, 1996; HARTMANN, 2017). São consideradas ancestrais evolutivos devido suas características estruturais e funcionais como a presença de regiões conservadas (CERRITELLI; CROUCH, 2008; MA, et al., 2008).

O quarto maior clã no PFAM encontrado no *cluster* nº 1, denominado Superfamília das Nucleosídeo Trifosfato Hidrolases (*P-loop containing nucleoside triphosphate hydrolase superfamily*) contém 229 Famílias e 2.787.646 domínios, e inclui 4 das 6 principais classes de enzimas (WEBB, 1992). As proteínas pertencentes a esta superfamília apresentam padrões conservados em suas sequências, como os motivos Walker A e Walker B (WALKER, et al, 1982) e 21 funções diferentes de acordo com os termos do identificador E. C (KAWAMURA, et al., 2003), com membros que podem atuar como quinases, proteínas motoras e

chaperonas (JIANPING; WEI; XIANWU, 2006).

O clã Superfamília dobra de Rossman ou domínio de Rossman (*FAD/NAD(P)-binding Rossmann fold Superfamily*, em inglês) é o quinto maior clã encontrado no *cluster* nº 1. Este grupo é formado por enzimas oxidorredutases que possuem o motivo estrutural dobra de Rossman (RAO; ROSSMAN, 1973). Este motivo está presente em inúmeras proteínas, sendo encontrado principalmente em cofatores que se ligam a nucleotídeos como FAD e NAD (CAETANO-ANOLLÉS; KIM; MITTENHAL, 2007; HANUKOGLU, 2015; MA, et al., 2008).

Os demais 261 clãs presentes no *cluster* nº 1 (APÊNDICE 7) apresentaram uma quantidade bem menor de sequências classificadas e por este motivo não foram detalhados nesta seção. Entretanto, analisando os clãs presentes no *cluster* nº1 verificamos que este é um *cluster* que apresenta inúmeros membros com características de proteínas muito antigas que já foram descritas na literatura (CAETANO-ANOLLES; CAETANO-ANOLLES, 2003; CAETANO-ANOLLES; CAETANO-ANOLLES, 2005), como por exemplo as dobras *P-loop containing nucleoside triphosphate hidrolases* (CL0023), *TIM beta/alfa barrel* (CL0036), *NAD(P)-binding Rossman-fold* (CL0063), *ferredoxin-like* (CL0344), *flavodoxin-like* (CL0325) e motivo *Ribonuclease H-like* (CL0219). Este achado pode sugerir um possível ancestral comum entre as proteínas deste cluster e também é um indicativo de que o valor de corte de 50% de identidade é suficiente para encontrar relações de homologia conforme descrito no trabalho de Pearson (2005).

6.2.1.2 Famílias

Neste estudo de caso biológico, foram identificadas 1685 Famílias diferentes de acordo com os resultados obtidos utilizando o conjunto de dados Pfam-A.

A Família mais abundante encontrada no conjunto de dados foi oligo_HPY (*oligo_HPY*, pfam08352), seguido por Nucleotídil-transferase (*NTP_transferase*, pfam00483), HIV-1 (*Vif*, pfam00559), proteína Gid-A (*GIDA*, pfam01134) e Actina (*Actin*, pfam00022), conforme pode ser visto na FIGURA 10.

A Família denominada *oligo_HPY* é formada por proteínas que apresentam uma sequência de aminoácidos específica encontrada na região C terminal dos peptídeos do tipo ABC, logo após o domínio de ligação ao ATP (*pfam00005*) (OLDEHINKEL; DOEVEN; POOLMAN, 2006). O domínio de ligação ao ATP foi um

dos mais abundantes dentre os 1377 encontrados no cluster n° 1, conforme apresentado na próxima seção.

A segunda maior Família encontrada, denominada Nucleotídil-transferase é composta por várias enzimas que transferem nucleotídeos para fosfo-açucares, como por exemplo as DNA polimerases que participam de processos como o reparo do DNA (ARAVIND; KOONIN, 1999; BATRA, et al., 2013; BEARD; WILSON, 2014).

A terceira maior Família é a das proteínas Vif do vírus da imunodeficiência humana (HIV). A proteína vif, também conhecida como fator de virulência, é a responsável por aumentar a infectividade do vírus participando apenas de uma etapa específica no ciclo do HIV. Ela se liga a proteína APOBEC3G e a degrada para permitir a replicação do vírus (LETKO., et al, 2015; ROSE., et al, 2004). O fato desta proteína compartilhar no mínimo 50% de identidade com as mais diversas proteínas que compõem o cluster n° 1 pode ser um indício de que esta proteína do HIV compartilha características estruturais ou funcionais inusitadas com as demais proteínas que compõem o *cluster*.

A Família das proteínas GidA (do inglês *Glucose inhibited Division protein A*) ou MnmG é a quarta maior encontrada na análise do *PfamScan*. Embora a literatura descreva esta proteína como conservada entre vários procariotos, a sua função ainda não foi totalmente esclarecida (WHITE, et al., 2008), porém considera-se que elas atuam juntamente com as MnmE na modificação de RNAs transportadores (FISLAGE; WAUTERS; VERSÉES, 2016), além de participar da divisão celular e replicação em bactérias e também na mitocôndria. Sua disfunção e variações a nível de sequência estão relacionadas a doenças mitocondriais (SHA, et al., 2004; WHITE, et al., 2008).

A quinta maior Família é a das Actinas. Esta Família é formada por proteínas que exercem múltiplas funções, como contração muscular, sinalização celular, motilidade, divisão celular e manutenção de junções celulares, além de formar os microtúbulos, sendo, portanto, o maior componente do citoesqueleto (DOHERTY; McMAHON, 2007; HERMAN, 1993; KABSCH; VANDERKERCKHOVE, 1992). A actina é uma proteína muito conservada entre diversas espécies de Eucariotos, pois permaneceu substancialmente inalterada durante bilhões de anos que separaram leveduras e humanos (ERICKSON, 2007; GUNNING, et al., 2009-2019).

Assim como na categoria de Clãs, a análise e comparação dos resultados na categoria de famílias do Pfam nos mostra uma grande quantidade de grupos de

proteínas com características conservadas entre diferentes espécies e que de acordo com os trabalhos de CAETANO-ANOLLÉS e col. (2007) e TODD e col. (2001) são funções de proteínas antigas e conservadas.

6.2.1.3 Domínios

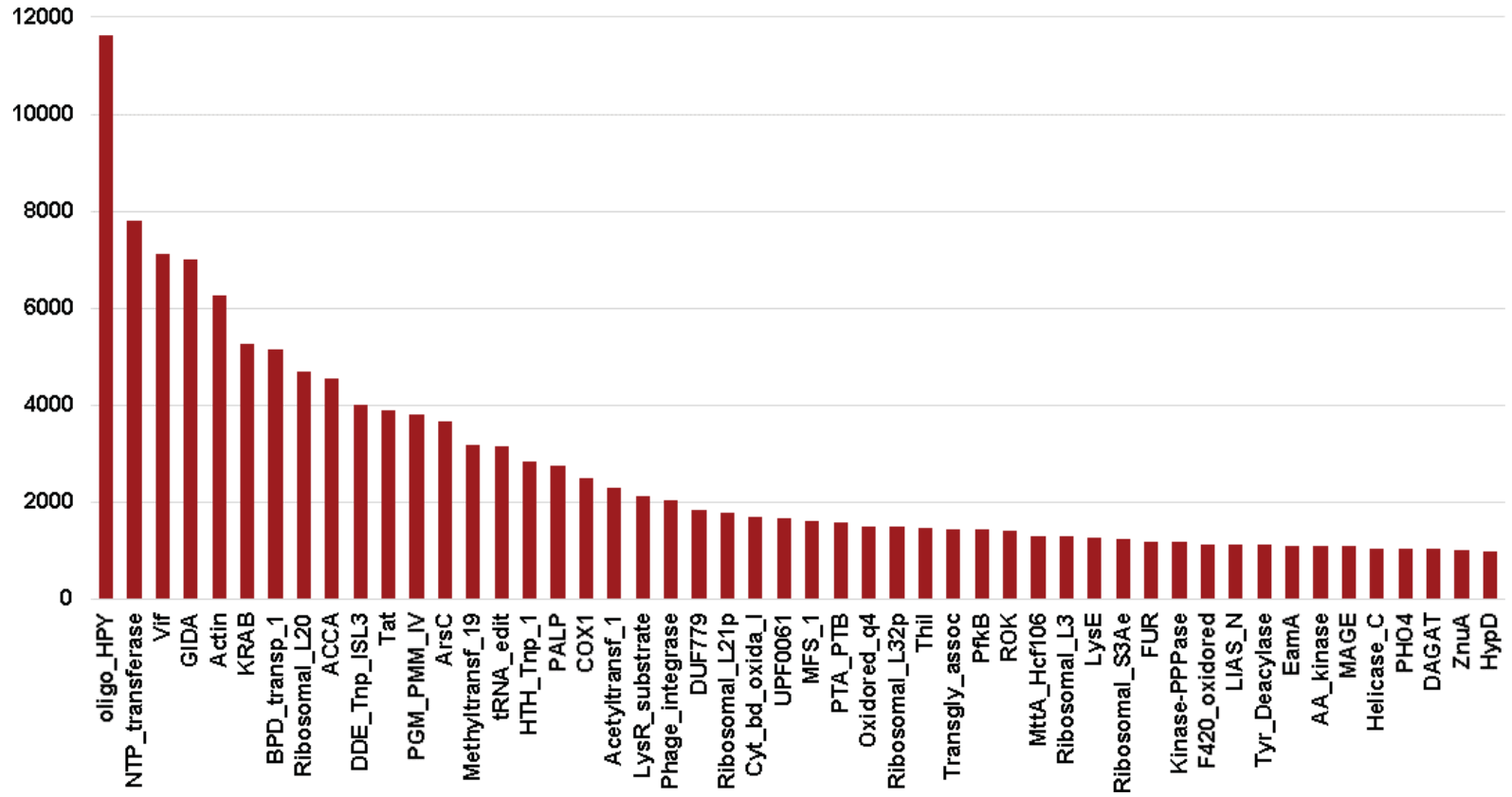
Foram identificados 1377 domínios proteicos nesta análise de acordo com os resultados obtidos utilizando o conjunto de dados Pfam-A. Os três domínios em maior abundância são os 3 domínios estruturais que compõem o fator de elongação Tu (KRAB; PARMEGGIANI, 2002; NISSEN, et al., 1995; SERGIEV; BOGDANOV; DONTSOVA, 2005) (*pfam00009, pfam03144, pfam03143*), seguido pelo domínio de ligação ao ATP dos transportadores ABC (*pfam00005*), dedo de zinco (*pfam00096*), domínio variável das imunoglobulinas (*pfam07686*), ClpP (*pfam00574*) e domínio regulador de resposta (*pfam00072*).

A Família dos fatores de elongação de ligação ao GTP é formada por fatores de elongação, como por exemplos os fatores eEF-1 em eucariotos e EF-Tu em procariotos. O fator EF-Tu, também conhecido como EF-1α, atua fornecendo a energia necessária para ligação do RNAt ao sítio A do ribossomo durante a tradução (MOLLER; SCHIPPER; AMONS, 1987; NISSEN, et al., 1995). Os domínios de ligação ao GTP do fator de elongação Tu são regiões conservadas e comuns em proteínas dependentes de GTP que ligam o RNAt ao ribossomo (MERRICK; CAVALLIUS; KINZY, 1993). Os outros dois domínios estruturais apresentam conformação estrutural barril beta (NISSEN, et al., 1995; WANG, et al., 1997) e participam da ligação ao RNAt carregado. Uma característica curiosa destes dois domínios é a baixa identidade entre si mesmo sendo participantes do mesmo processo molecular.

O quarto maior grupo de domínios encontrado é composto por dois dos quatro domínios proteicos que compõem os transportadores do tipo ABC (do inglês *ATP-binding cassette*). Os transportadores ABC são um grupo de proteínas de membrana presentes em mamíferos, fungos e procariontes, sendo os responsáveis pela hidrólise de ATP e fornecimento de energia para vários processos moleculares, além de transportar diversas moléculas e íons (DEAN; RZHETSKY; ALLIKMETS, 2001). Os domínios descritos nesta entrada do PFAM são altamente conservados e apresentam motivos como Walker A e B e *loop H like* (VASILIOU; VASILIOU;

NEBERT, 2009).

FIGURA 11 - ANOTAÇÃO DO CLUSTER 1 POR MEIO DA CLASSIFICAÇÃO DAS FAMÍLIAS DE HOMOLOGOS DO PFAM



FONTE: A Autora (2018).

LEGENDA: Representação gráfica da distribuição das sequências nas 50 maiores Famílias de Homólogos encontradas no Cluster nº 1 de acordo com as classificações do banco de dados PFAM.

O quinto maior grupo de domínios encontrados é formado pelos domínios dedo de zinco do tipo C2H2, ou seja, contém duas cisteínas e duas histidinas. Este domínio é um dos mais numerosos em genomas de eucariotos. Além de apresentar estruturas tridimensionais variadas, eles exercem diferentes funções, como por exemplo, ativação de transcrição, regulação da apoptose, dobramento de proteínas e reconhecimento de moléculas de DNA. Existe muitos tipos de dedo de zinco, porém os mais comuns nos bancos de dados são C2H2, CCHC e PHD (LAITY; LEE; WRIGHT, 2001; WOLFE; NEKLUDOVA; PABO, 2000).

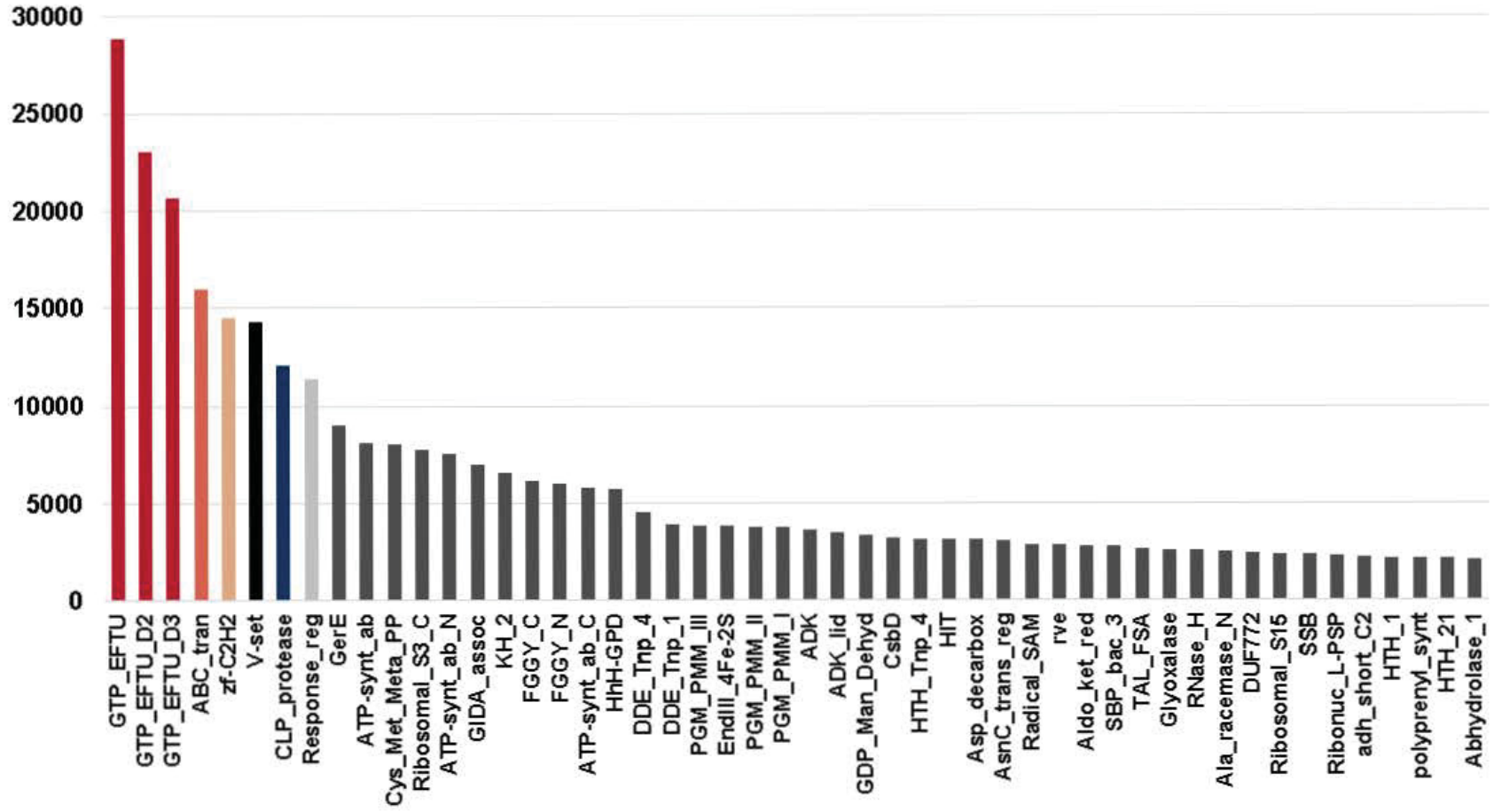
O domínio V-set é o sexto maior grupo encontrado nos resultados obtidos na análise realizada. Os domínios V-set são parecidos com os domínios das cadeias variáveis dos anticorpos. São encontrados em várias proteínas, como por exemplo, receptores de células T e moléculas de junção (KIM, et al., 2010).

O sétimo maior grupo de domínios encontrado é o das subunidades proteolíticas das proteases Clp. Clp é uma protease dependente de ATP responsável pela clivagem de proteínas, remoção de proteínas disfuncionais e controle do crescimento celular. Estas proteases são conservadas entre procariotos e eucariotos (ANDERSSON, et al.,2009; MAURIZI, et al., 1990).

Os domínios receptores das proteínas reguladoras de resposta (RR) formam o oitavo maior grupo de domínios encontrados. As proteínas reguladoras de resposta (RR) fazem parte do sistema regulador de dois componentes em bactérias e são definidas pela presença deste domínio (BOURRET, 2010).

Os demais domínios encontrados não foram detalhados nesta seção devido a menor quantidade de sequências classificadas. Porém, estes dados podem ser acessados em: https://github.com/grazLet/DadosDisserta-o

FIGURA 11 - ANOTAÇÃO DO CLUSTER 1 POR MEIO DA CLASSIFICAÇÃO DE DOMÍNIOS DO PFAM



FONTE: A Autora (2018).

LEGENDA: Domínios do PFAM identificados no *cluster* nº 1. Em pink os três domínios mais abundantes no conjunto de dados (fator de elongação Tu).

6.2.1.4 Unidades Polipeptídicas de Repetição

Foram identificadas 55 unidades polipeptídicas de repetição diferentes no *cluster* nº 1 de acordo com os resultados obtidos na análise realizada utilizando o conjunto de dados Pfam-A (FIGURA 12) .

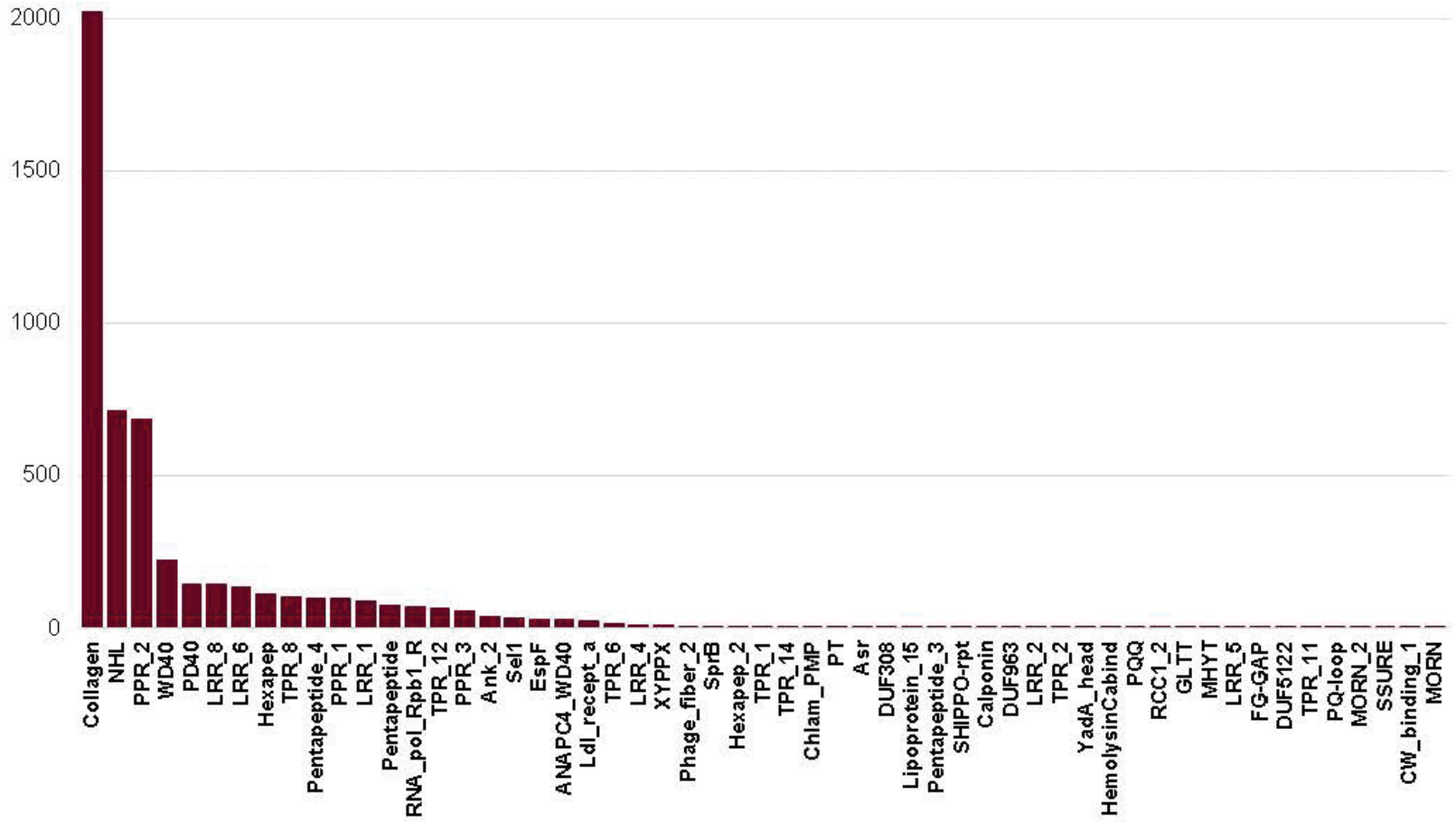
A unidade polipeptídica de repetição mais abundante nos resultados é a da tripla hélice do colágeno (*pfam01391*), seguida pela NHL (*pfam01436*) e pentatricopeptídeo (*pfam13041*).

Uma grande quantidade de sequências foi classificada como tendo a tripla hélice do colágeno. Esta região apresenta uma conformação com padrão limitado que requer a sequência repetitiva Glicina – X – Y em que geralmente X e Y são uma prolina e uma hidroxiprolina, respectivamente. A Família do colágeno mostra que este é um motivo adaptável a diferentes proteínas e funções, podendo ser essencial devido sua capacidade de se auto-associar a estruturas supramoleculares e também de interagir com ligantes e receptores (BRODSKY; PERSIKOV, 2005; MAINE; BREWTON, 1993).

A segunda unidade polipeptídica de repetição mais abundante no *cluster* n° 1, NHL, também conhecida como NCL-1, HT2A ou Lin-41 é uma sequência de aminoácidos encontrada em muitas proteínas de procariotos e eucariotos, como por exemplo, serinas quinases e mono-oxigenases (HUSTEN; EIPPER, 1991; SLACK; RUVKUN, 1998).

A terceira unidade polipeptídica de repetição mais numerosa, denominada pentatricopeptídeo são sequências de 35 aminoácidos que regulam a expressão gênica se ligando a uma região específica do RNAt. As proteínas que possuem essa repetição formam a maior família de proteínas mediadoras do controle pós-transcricional em organelas (MANNA, 2015).

FIGURA 12 - UNIDADES POLIPEPTÍDICAS DE REPETIÇÃO IDENTIFICADAS NO CLUSTER 1 DE ACORDO COM O BANCO DE DADOS PFAM



FONTE: A Autora (2018).

6.2.1.5 Coiled-coils

Foram identificadas 8 regiões do tipo *coiled-coil* no cluster nº 1 de acordo com os resultados obtidos na análise realizada utilizando o conjunto de dados Pfam-A, conforme apresentado na FIGURA 13.

Dentre as regiões identificadas duas se destacam pelo grande número de sequências: tropomiosina (*pfam00261*) e prefoldina (*pfam02996*).

A tropomiosina é a proteína que regula a contração muscular e esquelética junto a troponina. Ela interage com a actina dando sustentação durante o processo de contração muscular, exercendo assim um papel fundamental para o bom funcionamento da actina (XIAOCHUAN, et al., 2010; NITANAI, et al., 2010).

A prefoldina também é uma proteína cuja função também está relacionada a actina. Trata-se de uma cochaperona altamente especializada presente em todos os eucariotos e arqueias, que atua durante o dobramento dos filamentos de actina e tubulina durante a montagem do citoesqueleto (MARTIN-BENITO, et al., 2002; MILLAN-ZAMBRANO; CHÁVEZ, 2014).

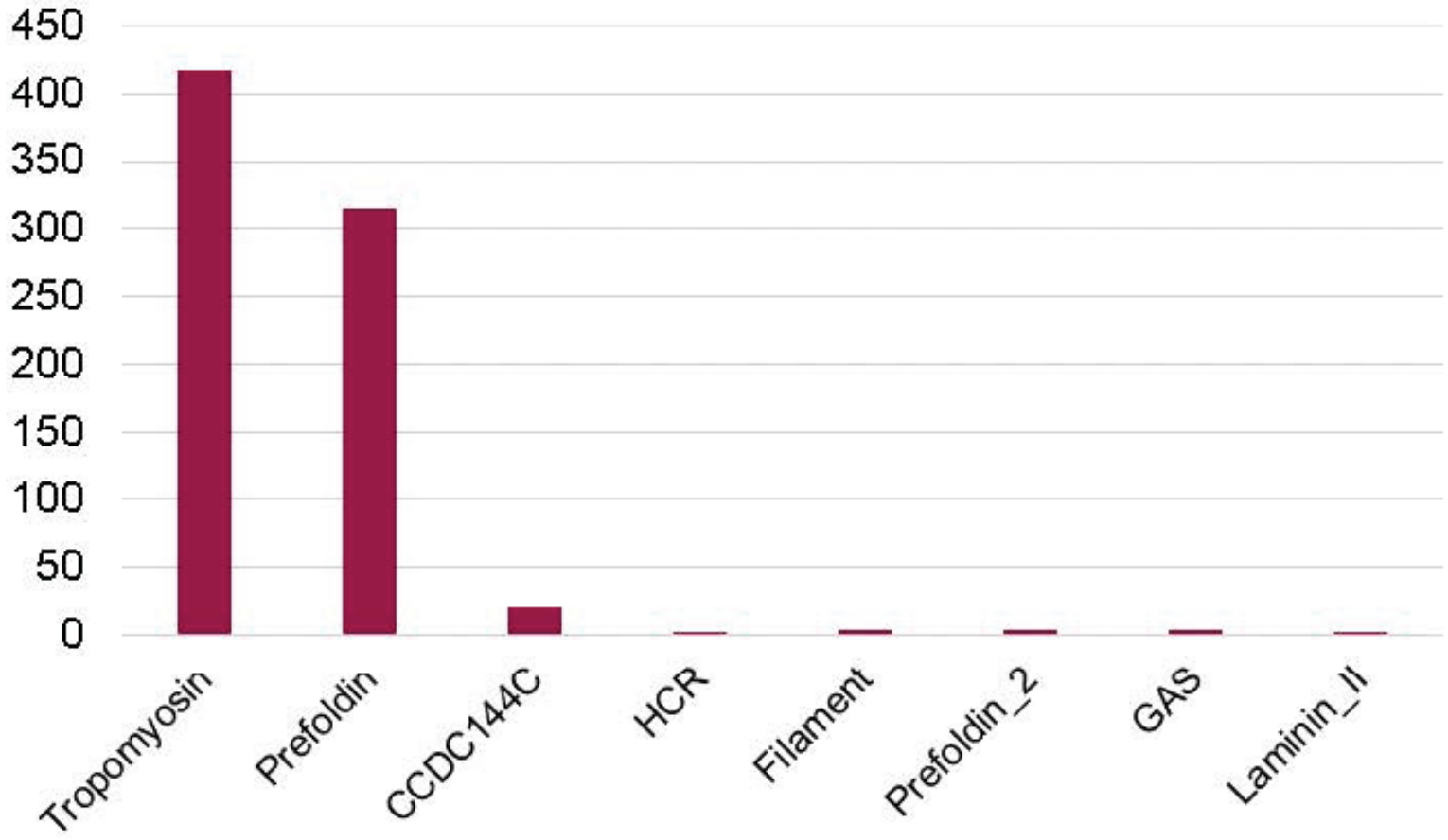
6.2.1.6 Regiões Desordenadas

A análise realizada utilizando o conjunto de dados Pfam-A identificou apenas a região desordenada cornifina (*pfam02389*) em uma sequência parcial de antígeno de merozoíto.

TABELA 2 - REGIÃO DESORDENADA IDENTIFICADA NO CLUSTER 1 DE ACORDO COM O BANCO DE DADOS PFAM

NCBI ID	**PFAM ID**	***Bit-score***
gi|294955990|ref|XP_002788777.1|	*Cornifin*	26,9

FIGURA 13 -REGIÕES COILED-COILD IDENTIFICADAS NO *CLUSTER* 1 DE ACORDO COM O BANCO DE DADOS PFAM



FONTE: A Autora (2018).

6.3 Caracterização Funcional Baseada nas Anotações do *Gene Ontology*

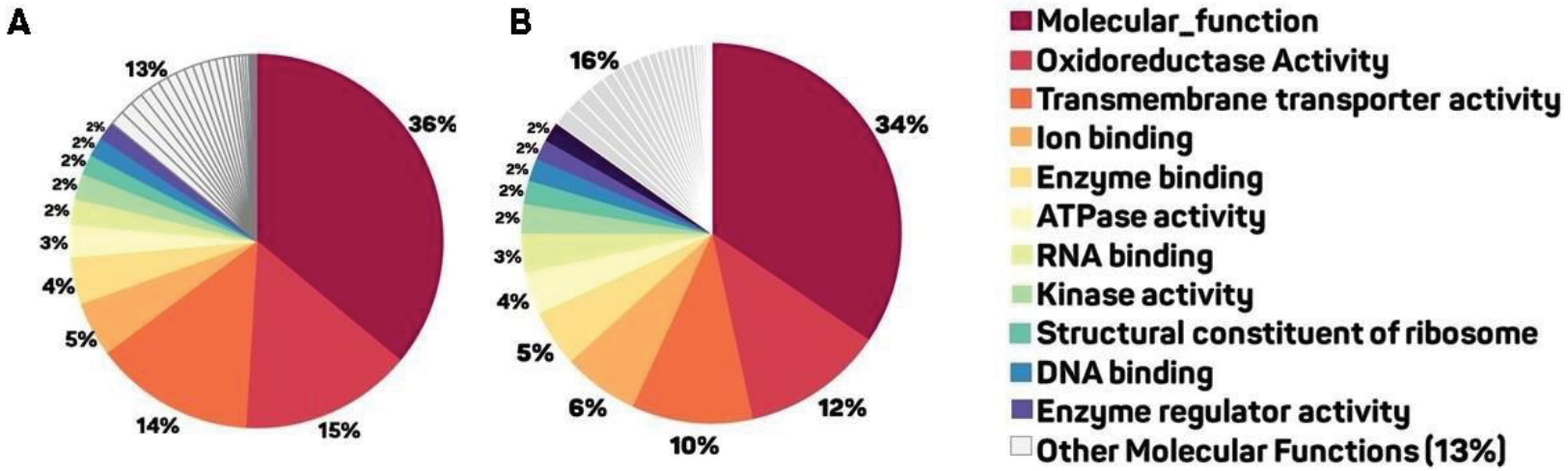
Para caracterizar as propriedades funcionais dos 21 maiores clusters obtidos foi realizado o enriquecimento funcional por meio dos dados disponíveis no banco *Gene Ontology* (ASHBURNER et. al., 2000) e também a avaliação da consistência *intracluster* com base nos resultados obtidos. É importante ressaltar que a consistência entre os termos GO obtidos pode variar em diferentes níveis, indo desde muitas sequências do mesmo cluster compartilhando termos GO idênticos até membros com termos GO relacionados apenas pelo termo raiz da ontologia, como por exemplo *Molecular Function* (GO: 0003674), *Cellular Component* (GO:0005575) e *Biological Process* (GO:0008150).

6.3.1 Caracterização do Cluster Número 1

Após a obtenção das anotações do GO para todas as sequências do *cluster* n° 1, estes resultados foram sumarizados e gráficos de pizza foram produzidos para representar a distribuição das anotações nas 3 categorias do GO: *Molecular Function*, *Cellular Component* e *Biological Process*.

Na categoria *Molecular Function* (FIGURA 14), tanto nas análises utilizando os bancos de dados UniProtKB (The UniProt Consortium, 2017) (FIGURA 14 A) quanto nas em que o SwissProt (The UniProt Consortium, 2017) foi aplicado (FIGURA 14 B) observou-se uma grande quantidade de funções celulares básicas como atividade de oxidorredutase (GO:0016491), transporte transmembrana (GO:0022857), ligação ao DNA (GO:0003677) e atividade de ATPase (GO:0016887), o que pode sugerir que este é um cluster de funções universais e essenciais para diferentes proteínas e organismos (Acevedo-Rocha, et al., 2013).

FIGURA 14 - RESULTADO DO ENRIQUECIMENTO COM AS ANOTAÇÕES DO GENE ONTOLOGY NA CATEGORIA *MOLECULAR FUNCTION* PARA O CLUSTER Nº 1

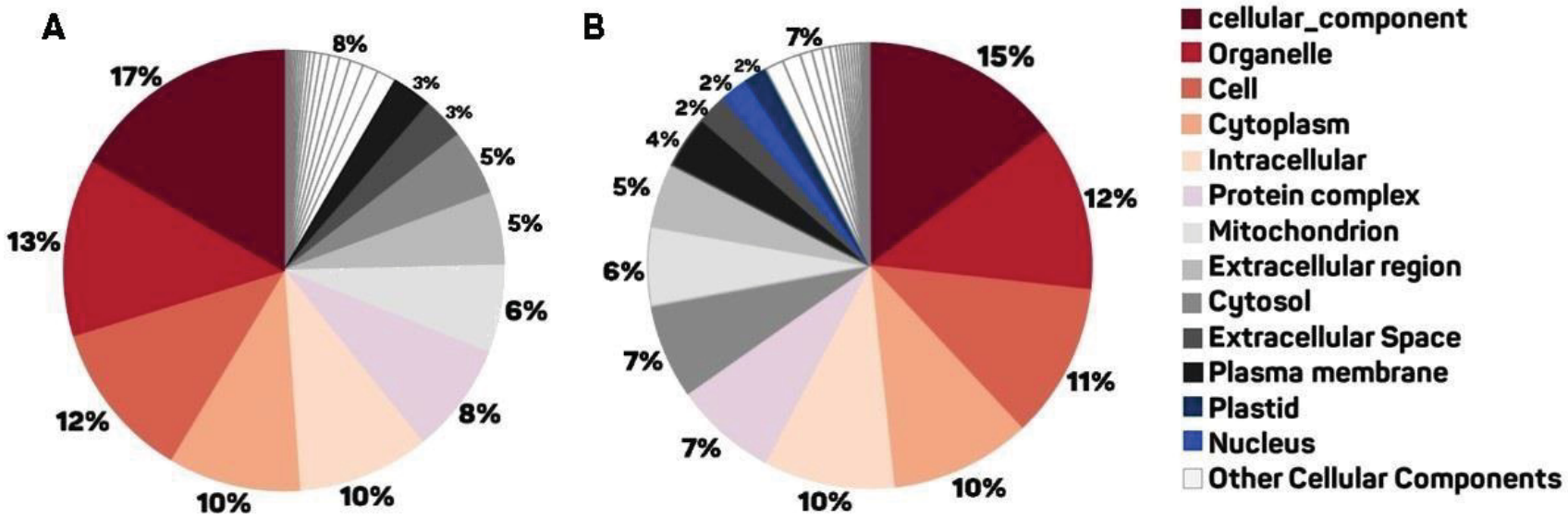


FONTE: A autora (2018).
LEGENDA: Gráficos de pizza mostrando a distribuição dos termos do *Gene Ontology* na categoria *Molecular Function*. (A) Resultado utilizando o banco de dados UniProtKB. (B) Resultado utilizando o banco SwissProt.

Na categoria *Cellular Component* (FIGURA 15), nas análises utilizando os bancos de dados UniProtKB (FIGURA 15 A) e SwissProt (FIGURA 15 B) observou-se que as sequências foram distribuídas em uma quantidade razoável de organelas celulares como mitocôndria (GO: 0005739), núcleo (GO: 0005634) e membrana plasmática (GO: 0044459), o que pode sugerir a heterogeneidade deste cluster antes do seu reprocessamento utilizando técnicas de mineração de dados específicas como a clusterização com alta porcentagem de identidade e/ou mineração de texto.

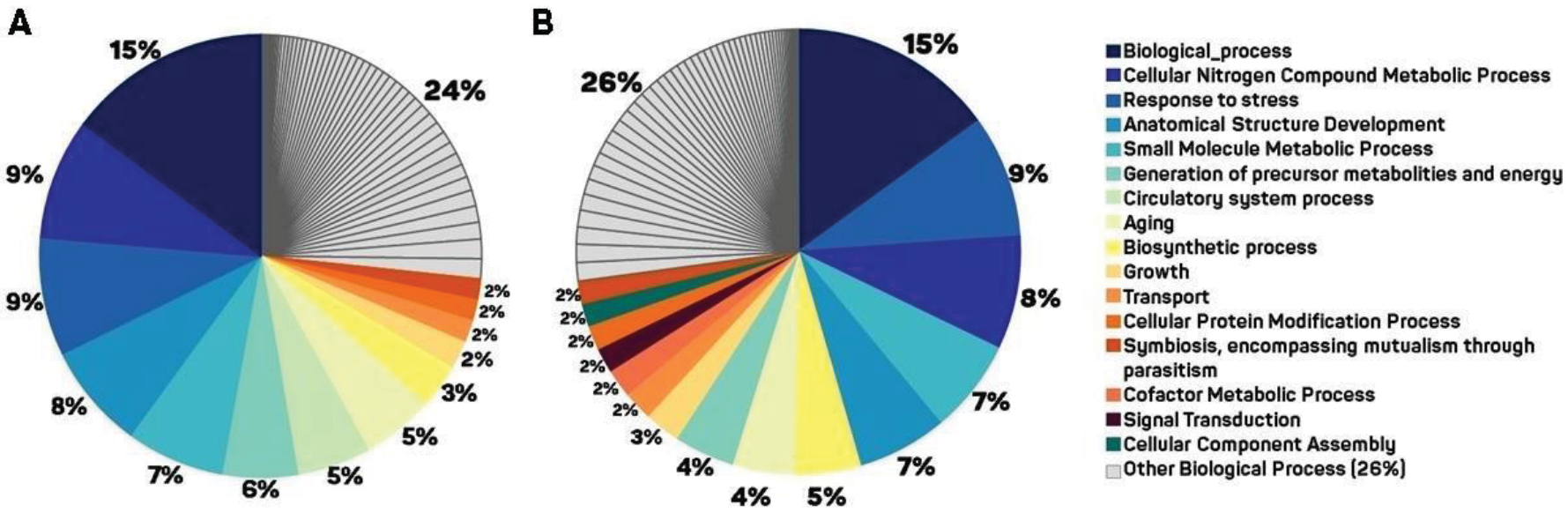
E finalmente analisando o conjunto de dados na categoria *Biological Process* (FIGURA 16) verificou-se que as sequências foram distribuídas em processos biológicos essenciais à grande maioria dos organismos como resposta ao estresse celular (GO: 0006950) e transporte (GO:0006810. Estas características foram observadas tanto nos resultados da análise que utilizou o banco de dados UniProtKB (FIGURA 16 A) quanto nos obtidos com o banco de dados SwissProt (FIGURA 16 B).

FIGURA 15 - RESULTADO DO ENRIQUECIMENTO COM AS ANOTAÇÕES DO GENE ONTOLOGY NA CATEGORIA CELLULAR COMPONENT PARA O *CLUSTER* Nº1.



FONTE: A Autora (2018).
LEGENDA: Gráficos de pizza mostrando a distribuição dos termos do Gene Ontology na categoria *Cellular Component*. (A) Resultado utilizando o banco de dados UniProtKB. (B) Resultado utilizando o banco SwissProt.

FIGURA *16* - RESULTADO DO ENRIQUECIMENTO COM AS ANOTAÇÕES DO GENE ONTOLOGY NA CATEGORIA BIOLOGICAL PROCESS PARA O *CLUSTER* Nº1.



FONTE: A Autora (2018).
LEGENDA: Gráficos de pizza mostrando a distribuição dos termos do Gene Ontology na categoria *Biological Process*. (A) Resultado utilizando o banco de dados UniProtKB. (B) Resultado utilizando o banco SwissProt.

6.3.2 Caracterização Funcional dos 21 Maiores Clusters

A etapa de caracterização e anotação dos 21 maiores *clusters* do banco de dados NR foi desenvolvida em MATLAB® e PERL® após a utilização da ferramenta GOSlimViewer (McCARTHY, et al., 2006). Nesta análise os *clusters* com mais de cem mil sequências (TABELA 3) foram selecionados para comparação dos resultados obtidos na anotação com os termos do Gene Ontology.

6.3.2.1 Função Molecular

Os resultados gerados na ferramenta GOSlimViewer para os 21 *clusters* na categoria *Molecular Function* (GO:003674) utilizando o banco de dados UniProtKB mostram que entre 14% e 100% dos membros em um dado *cluster* foram anotados no Gene Ontology (TABELA 3) e apresentam termos no banco de dados GOSlim.

Ao compararmos estes resultados (FIGURA 17) verificamos que a maior parte dos membros foi enriquecido com o termo raiz da ontologia (GO: 0003674), exceto os *clusters* nº 7 e 11 que apresentaram funções moleculares básicas como atividade de oxidorredutase (GO:0016491), transporte transmembrana (GO:0022857) e ligação ao DNA (GO:0003677) para a maior parte dos seus membros, como pode ser visto nas FIGURAS 18 e 19. Esta última observação e a predominância de determinados termos nos *clusters* nos permitiu inferir que todos os maiores *clusters* criados na clusterização do banco de dados NR são compostos por membros que apresentam funções moleculares básicas (FIGURAS 18 e 19), seguindo o mesmo padrão de características universais do maior *cluster*. Dentre estas funções, além das citadas acima temos: atividade de ATPase (GO:0016887), ligação a íons (GO:0043167), componente estrutural ribossomal (GO:0003735) e ligação a enzimas (GO:0019899).

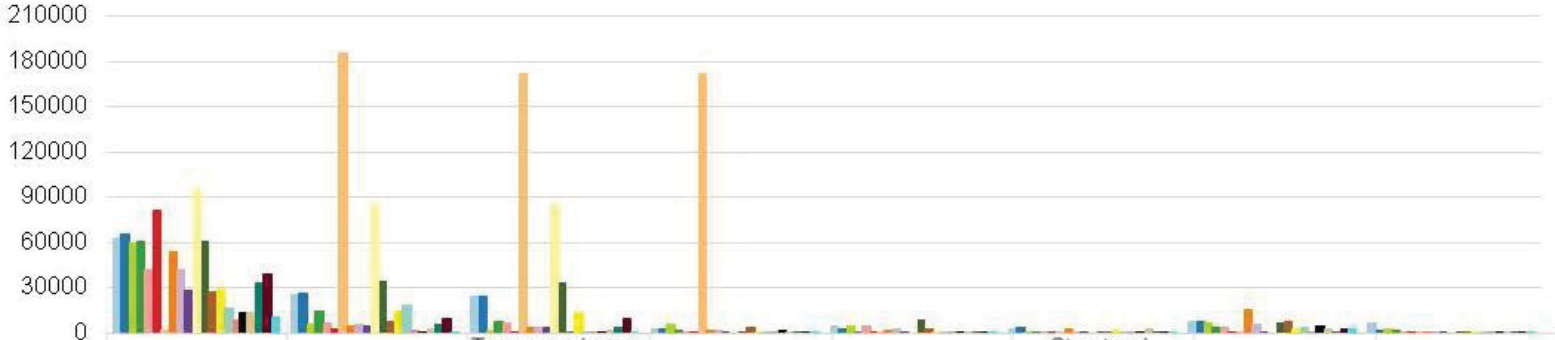
TABELA 3 - ESTATÍSTICAS DE TERMOS GO ANOTADOS NOS 21 CLUSTERS NA CATEGORIA *MOLECULAR FUNCTION*

Cluster ID	Sequências com Anotação no GOSlim
1	173.114/593.749 (29%)
2	163.909/440.501 (28%)
3	121.352/392.014 (28%)
4	111.631/339.845 (28%)
5	89.473/310.285 (28%)
6	101.407/308.645 (33%)
7	301.245/301.245 (100%)
8	110.346/273.541 (40%)
9	86247/252023 (34%)
10	44.729/168.018 (27%)
11	163.338/163.338 (100%)
12	159.416/159.416 (100%)
13	68.890/154.888 (44%)
14	74.830/139.947 (53%)
15	54.124/134.209 (40%)
16	17.554/124.622 (14%)
17	34.302/117.462 (29%)
18	34.374/115.109 (30%)
19	51.965/113.186 (46%)
20	69.753/111.223 (63%)
21	24.981/105.778 (24%)

FONTE: A Autora (2018).

Analisando os resultados gerados para os mesmos conjuntos de dados com a ferramenta GOanna utilizando o banco de dados SwissProt, observamos que embora o SwissProt seja um banco manualmente curado por especialistas na área, este fato aparentemente não influenciou na anotação baseada em homologia realizada, pois a distribuição dos termos nos clusters foi praticamente a mesma, conforme apresentado nas FIGURAS 20 e 21. Uma exceção neste caso foi o cluster nº 19, que ao contrário dos demais não apresentou nenhum membro anotado com o termo *Structural constituent of ribosome* (GO:0003735).

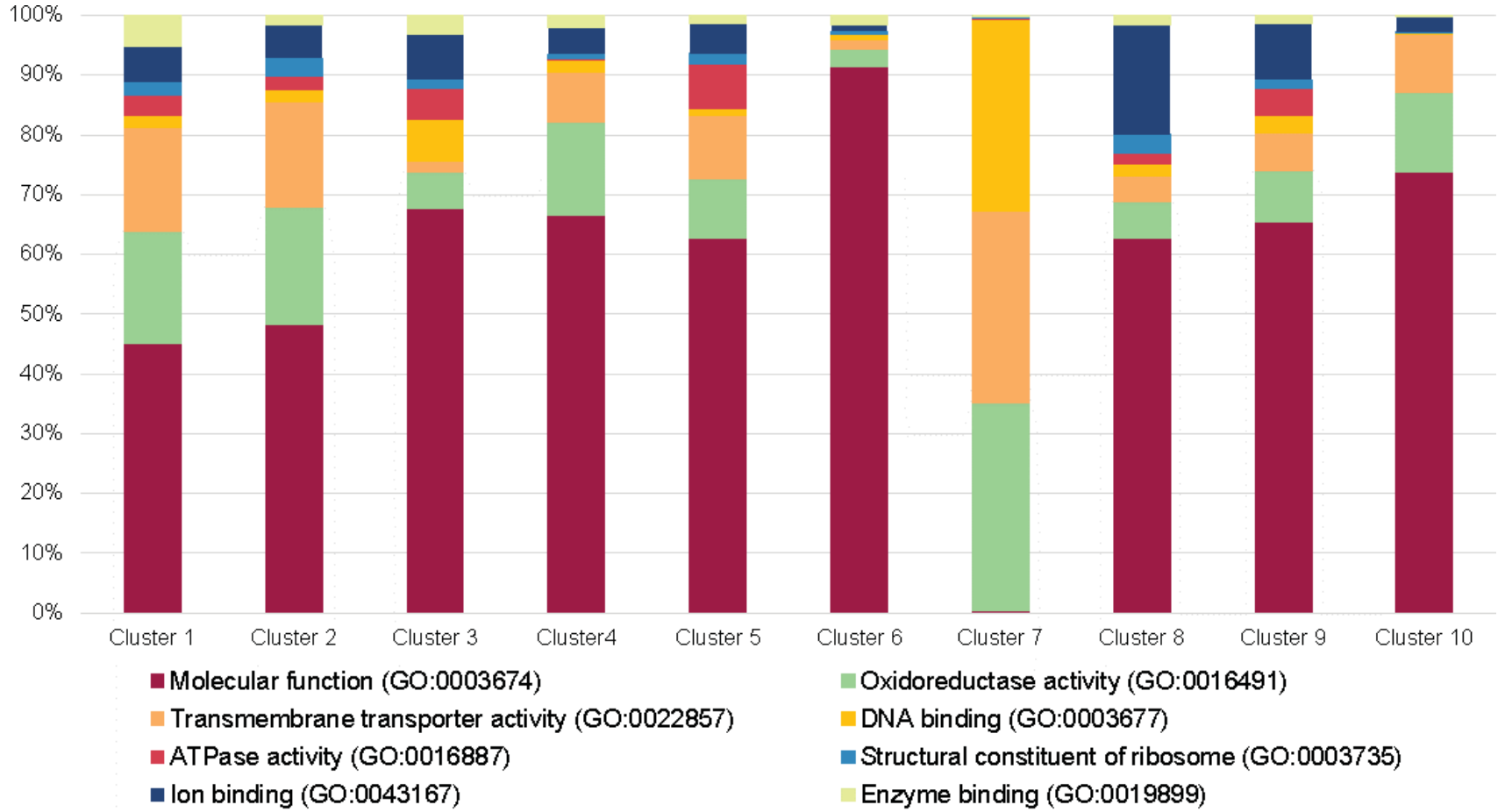
FIGURA 17 - GRÁFICO REPRESENTANDO A DISTRIBUIÇÃO DAS ANOTAÇÕES DOS MEMBROS DOS 21 CLUSTERS



	Molecular function (GO:0003674)	Oxidoreductase activity (GO:0016491)	Transmembrane transporter activity (GO:0022857)	DNA binding (GO:0003677)	ATPase activity (GO:0016887)	Structural constituent of ribosome (GO:0003735)	Ion binding (GO:0043167)	Enzyme binding (GO:0019899)
Cluster 1	62373	25831	24054	2925	4753	2989	8225	7051
Cluster 2	65615	26733	24186	2790	2793	4250	7689	1990
Cluster 3	59722	5403	1603	6078	4696	1337	6618	2754
Cluster4	61140	14395	7760	1783	317	583	4120	1843
Cluster 5	42538	6732	7197	712	5122	1140	3372	916
Cluster 6	81723	2606	1338	866	50	432	1011	1315
Cluster 7	1521	185461	171474	171387	888	1008	439	562
Cluster 8	54115	5288	3768	1786	1549	2608	15839	1342
Cluster 9	41974	5611	4012	1864	2904	936	6007	916
Cluster 10	28880	5202	3782	72	31	74	955	99
Cluster 11	95253	86027	85222	615	451	26	1380	353
Cluster 12	61467	34572	33629	56	9211	102	7619	592
Cluster 13	27550	7863	61	3777	2909	42	7610	686
Cluster 14	29615	14270	13416	697	141	1889	2663	1031
Cluster 15	16675	18974	1050	903	553	753	3724	781
Cluster 16	8894	1758	538	241	22	552	801	178
Cluster 17	14091	911	388	1856	741	296	4634	1357
Cluster 18	13429	3337	1760	470	357	2474	3093	540
Cluster 19	33030	5965	4210	75	33	94	1236	166
Cluster 20	38938	10069	9906	612	128	89	2727	232
Cluster 21	11089	351	1242	525	688	448	2372	142

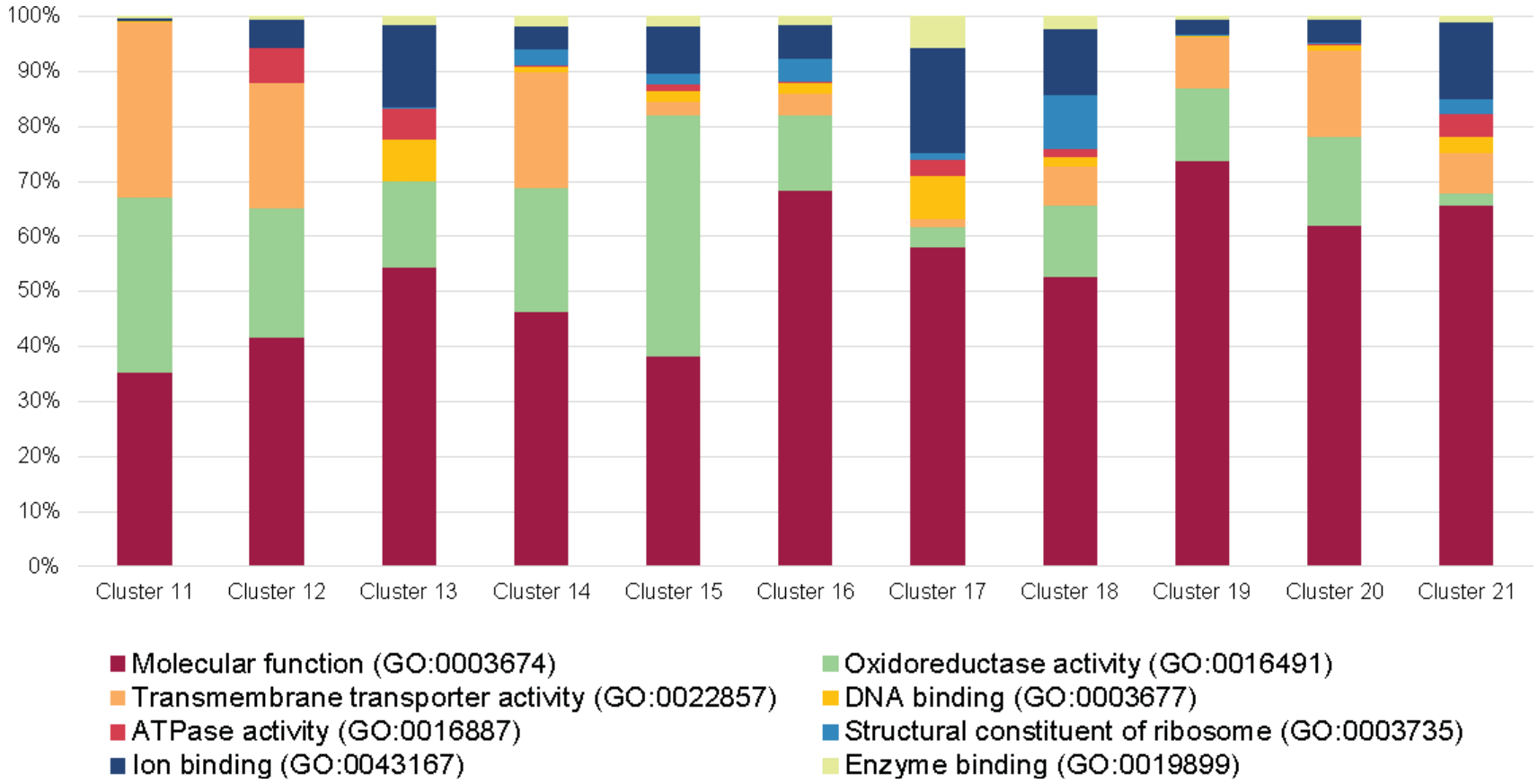
FONTE: A Autora (2018).
LEGENDA: Resultados da análise realizada com o banco de dados UNIPROTKB.

FIGURA 18 – DISTRIBUIÇÃO DOS TERMOS DO GENE ONTOLOGY NOS MAIORES CLUSTERS DO BANCO DE DADOS NR – PARTE A



FONTE: A Autora (2018).
LEGENDA: Resultados da análise realizada com o banco de dados UNIPROTKB. Neste gráfico temos a representação da distribuição dos termos GO que mais ocorrem nos 21 *clusters* na categoria *Molecular Function*.

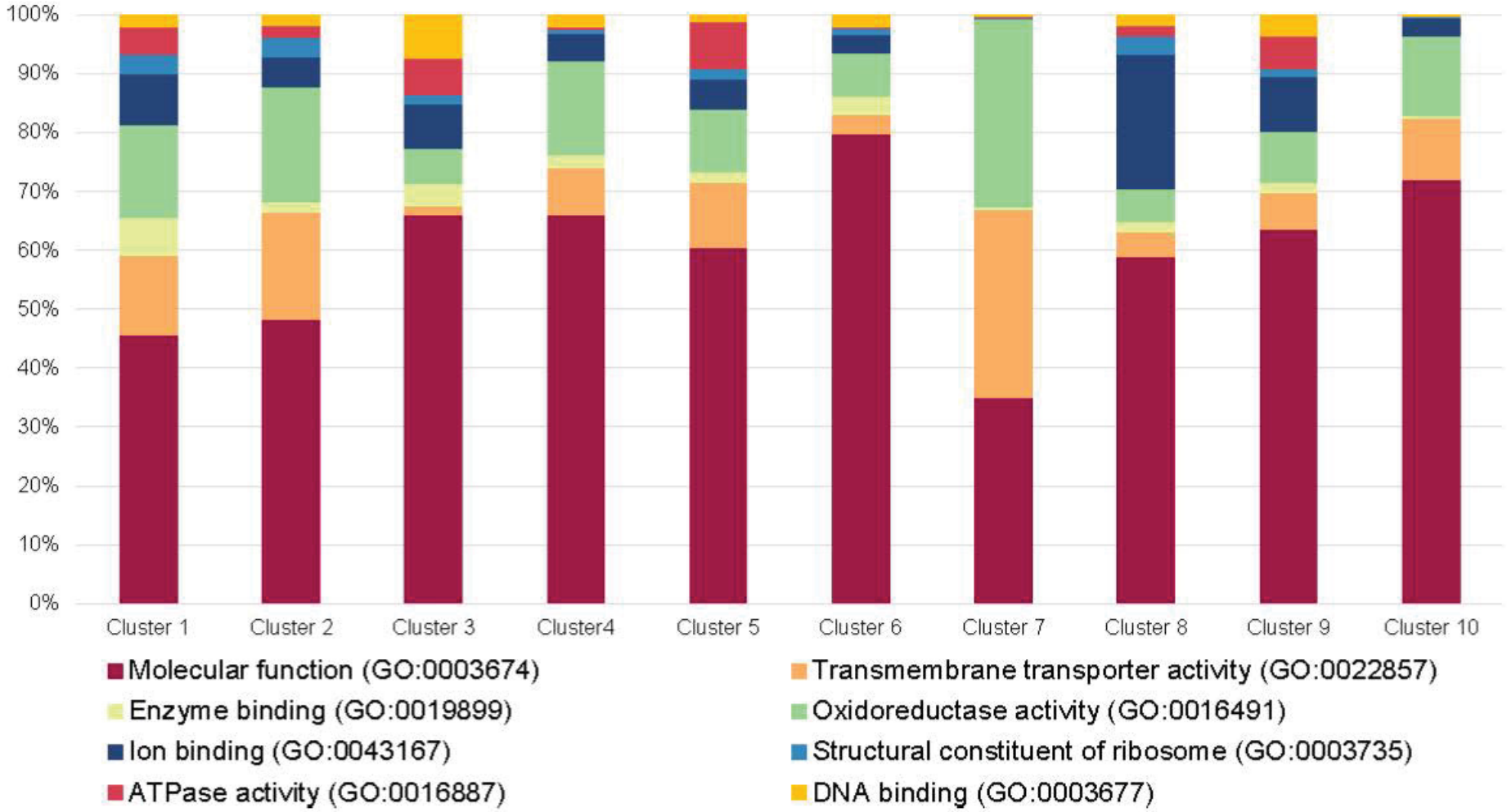
FIGURA 19 - DISTRIBUIÇÃO DOS TERMOS DO GENE ONTOLOGY NOS MAIORES CLUSTERS DO BANCO DE DADOS NR – PARTE B



FONTE: A Autora (2018).

LEGENDA: Resultados da análise realizada com o banco de dados UNIPROTKB. Neste gráfico temos a representação da distribuição dos termos GO que mais ocorrem nos 21 *clusters* na categoria *Biological Function*.

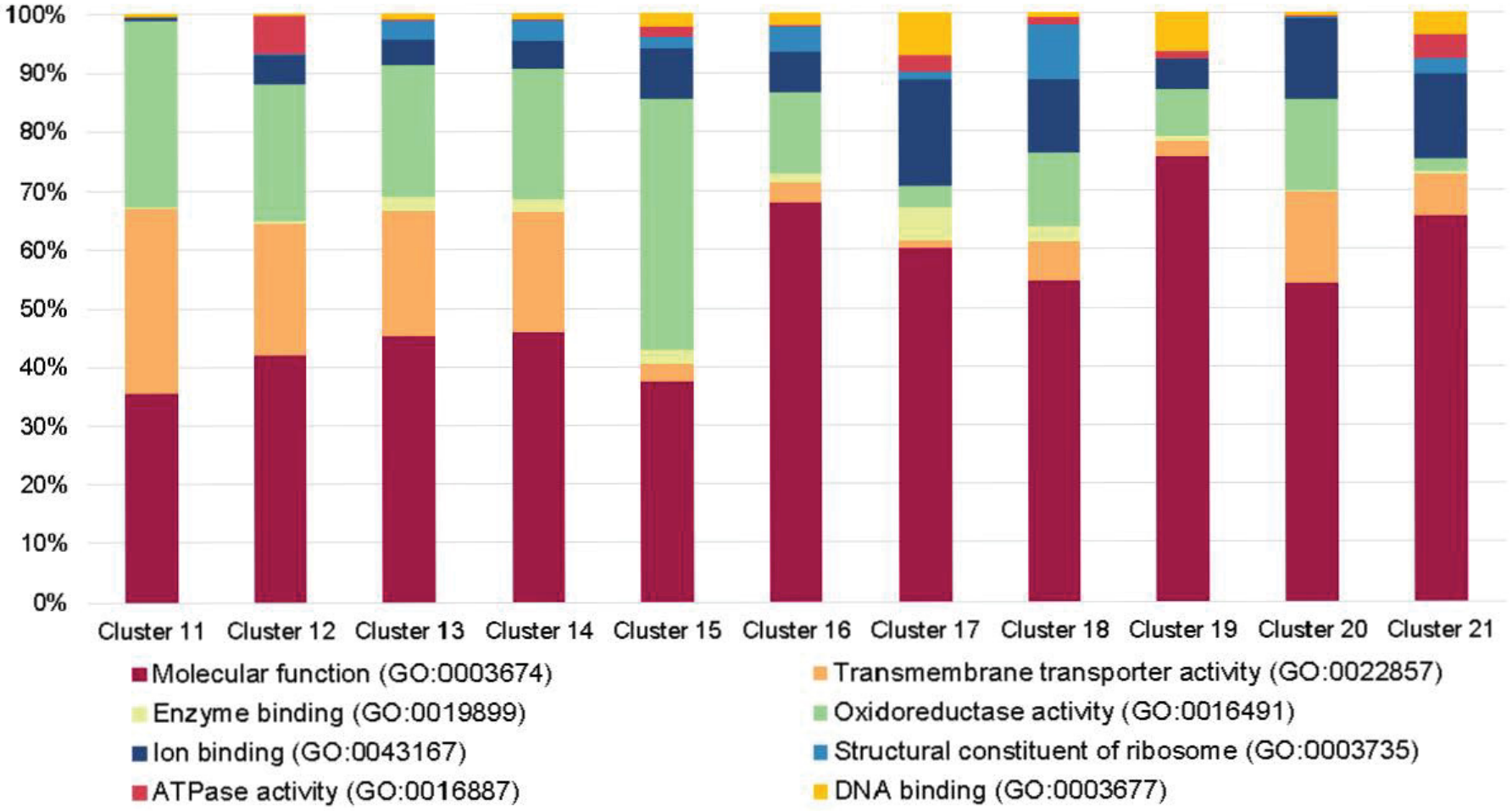
FIGURA 20 - DISTRIBUIÇÃO DOS TERMOS DO GENE ONTOLOGY NOS MAIORES CLUSTERS DO BANCO DE DADOS NR – PARTE C



FONTE: A Autora (2018).

LEGENDA: Resultados da análise realizada com o banco de dados SWISSPROT. Neste gráfico temos a representação da distribuição dos termos GO que mais ocorrem nos 21 *clusters* na categoria *Biological Function*.

FIGURA 21 - DISTRIBUIÇÃO DOS TERMOS DO GENE ONTOLOGY NOS MAIORES *CLUSTERS* DO BANCO DE DADOS NR – PARTE D



FONTE: A Autora (2018).
LEGENDA: LEGENDA: Resultados da análise realizada com o banco de dados SWISSPROT. Neste gráfico temos a representação da distribuição dos termos GO que mais ocorrem nos 21 *clusters* na categoria *Biological Function*.

6.3.2.2 Componente Celular

Embora o foco do estudo de caso biológico realizado neste trabalho fosse a caracterização funcional dos conjuntos de dados, os resultados obtidos na ferramenta GOSlimViewer utilizando os dados da análise realizada com os bancos de dados UniProtKB e SwissProt para a categoria *Cellular Component* (GO:0005575) também foram brevemente analisados com o intuito de encontrar algum padrão que fosse *cluster* específico.

A tabela 4 nos mostra que entre 29% e 100% dos membros em um dado *cluster* foram anotados nesta categoria e também no conjunto de dados Gene Ontology Slim, mostrando que aparentemente o Gene Ontology possui maior cobertura de anotações nesta categoria da ontologia.

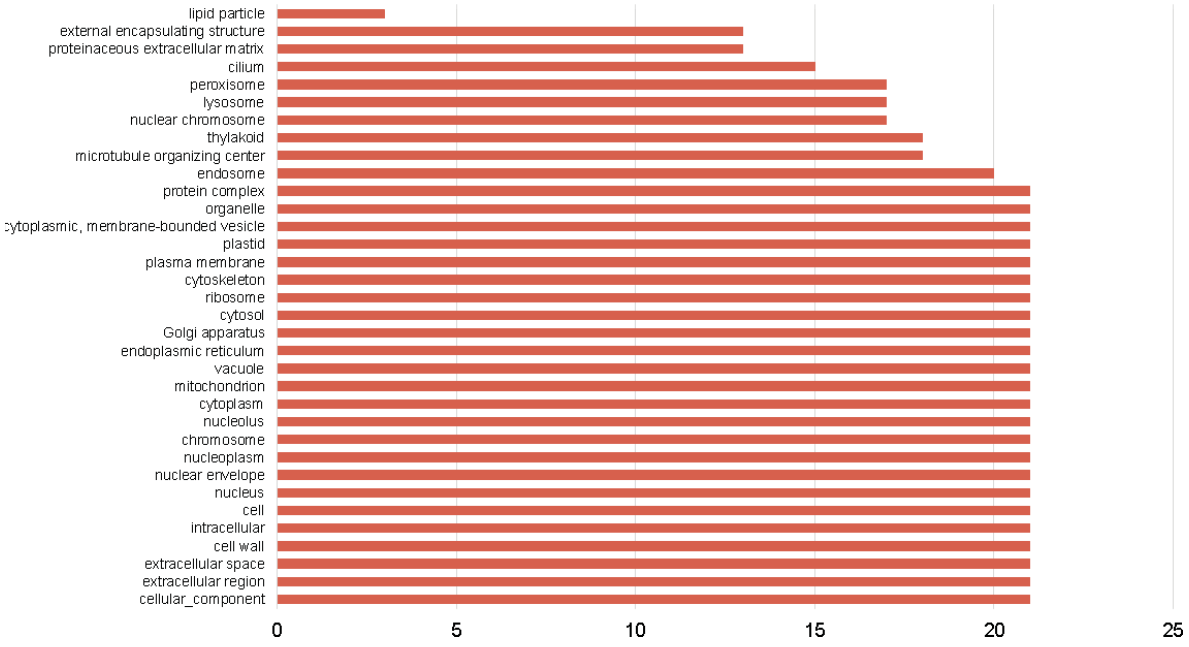
TABELA 4 - ESTATÍSTICA DE TERMOS GO ANOTADOS NOS 21 CLUSTERS NA CATEGORIA *CELLULAR COMPONENT*

Cluster ID	***Sequências com Anotação no GOSlim***
1	416.394/593.749 (70%)
2	316.856/440.501 (72%)
3	217.614/392.014 (56%)
4	250.095/339.845 (74%)
5	129.404/310.285 (42%)
6	105.957/308.645 (34%)
7	301.245/301.245 (100%)
8	156.096/273.541 (57%)
9	141.204/252.023 (56%)
10	161.127/168.018 (96%)
11	163.336/163.336 (100%)
12	159.416/159.416 (100%)
13	72.426/154.888 (47%)
14	102.497/139.947 (73%)
15	61.917/134.209 (46%)
16	36.335/124.622 (29%)
17	68.897/117.462 (59%)
18	70.771/115.109 (61%)
19	113.186/113.186 (100%)
20	72.121/111.223 (65%)
21	38.499/105.778 (36%)

FONTE: A autora (2018).

Ao compararmos os resultados obtidos para os 21 *clusters* na categoria *Cellular Component* observamos também que assim como na categoria *Molecular Function*, a maior parte dos membros foi enriquecido com o termo raiz da ontologia (GO: 0005575). E assim como no maior *cluster* do banco nós observamos que os termos GO se distribuem seguindo o mesmo padrão em todos os *clusters*, conforme pode ser visto na FIGURA 22. Portanto, chegamos à conclusão de que nesta categoria do GO não há nenhum padrão cluster específico dentre os conjuntos de dados estudados, o que também pode indicar a alta semelhança entre eles.

FIGURA 22 – OCORRÊNCIA DOS TERMOS GO (CATEGORIA CELLULAR COMPONENT) NOS 21 *CLUSTERS*



FONTE: A Autora (2018).

LEGENDA: No eixo x temos a quantidade de ocorrência dos termos GO nos 21 clusters e no eixo y temos o nome dos termos GO.

6.3.2.2 Processo Biológico

Os resultados gerados na ferramenta GOSlimViewer para os 21 *clusters* na categoria *Biological Process* (GO:003674) utilizando o banco de dados UniProtKB mostram que entre 27% e 100% dos membros em um dado *cluster* foram anotados no Gene Ontology (TABELA 5) e apresentam termos no banco de dados GOSlim.

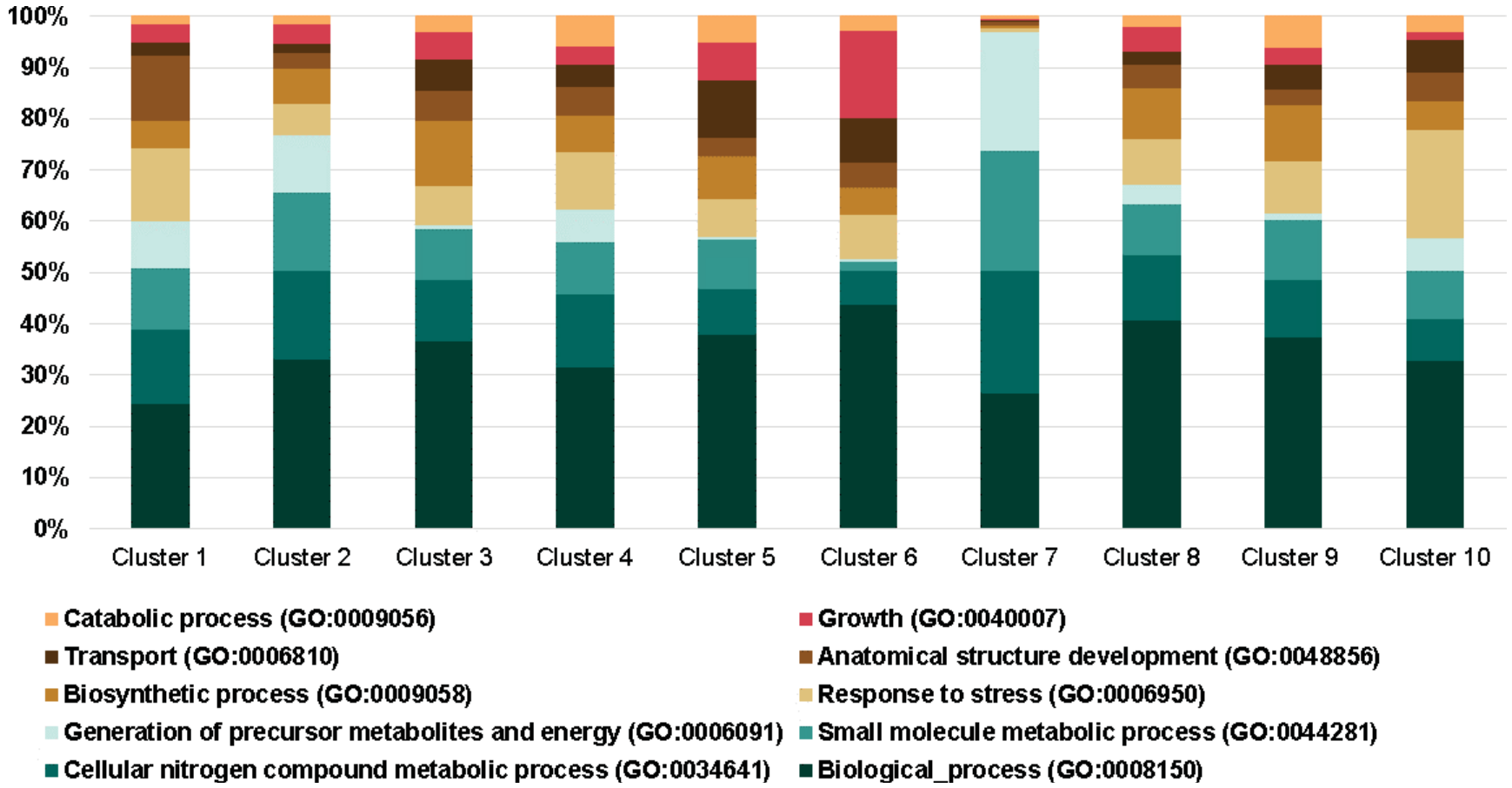
TABELA 5 - ESTATÍSTICAS DE TERMOS GO ANOTADOS NOS 21 CLUSTERS NA CATEGORIA *BIOLOGICAL PROCESS*

Cluster ID	Sequências com Anotação no GOSlim
1	439.374/593.749 (74%)
2	291.399/440.501 (66%)
3	282.603/392.014 (72%)
4	239.288/339.845 (70%)
5	163.596/310.285 (53%)
6	164.470/308.645 (50%)
7	301.245/301.245 (100%)
8	273.023/273.541 (54%)
9	158.078/252.023 (63%)
10	79.227/168.018 (47%)
11	163.338/136.338 (100%)
12	159.416/159.416 (100%)
13	109.000/154.888 (70%)
14	110.388/139.947 (79%)
15	70.009/134.209 (52%)
16	33.476/124.622 (27%)
17	82.695/117.462 (70%)
18	77.009/115.109 (67%)
19	94.478/113.186 (82%)
20	89.093/111.223 (80%)
21	46.351/105.778 (44%)

FONTE: A Autora (2018).

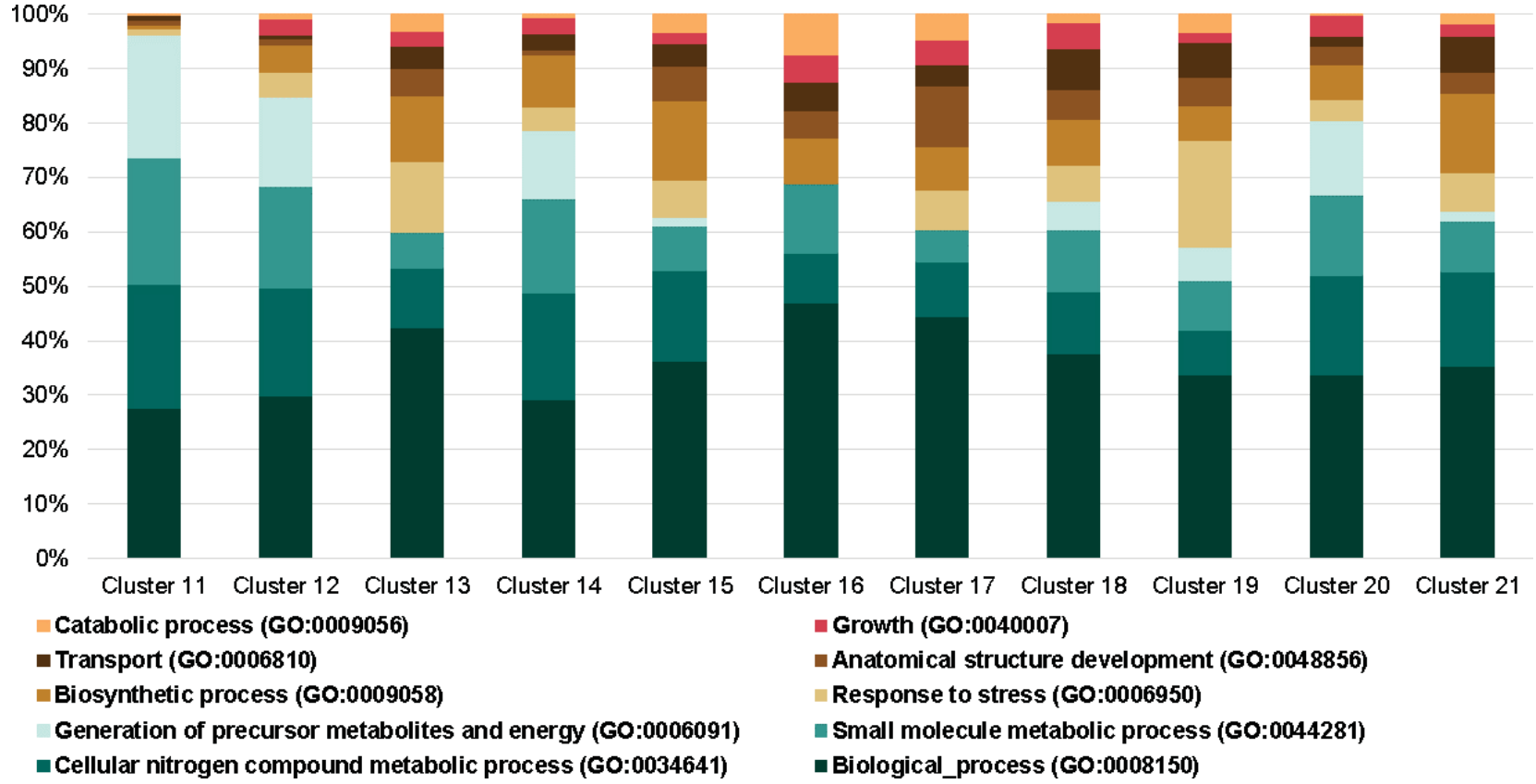
Ao compararmos estes resultados verificamos que assim como nas outras duas categorias do GO, a maior parte dos membros foi enriquecido com o termo raiz da ontologia (GO: 0008150), exceto os *clusters* nº 7 e 11, conforme pode ser visto nas FIGURAS 23 e 24.

FIGURA 23 - DISTRIBUIÇÃO DOS TERMOS DO GENE ONTOLOGY NOS MAIORES CLUSTERS DO BANCO DE DADOS NR – PARTE A



FONTE: A Autora (2018).
LEGENDA: Resultados da análise realizada com o banco de dados UNIPROTKB. Neste gráfico temos a representação da distribuição dos termos GO que mais ocorrem nos 21 *clusters* na categoria *Biological Process*

FIGURA 24 - DISTRIBUIÇÃO DOS TERMOS DO GENE ONTOLOGY NOS MAIORES CLUSTERS DO BANCO DE DADOS NR – PARTE B



FONTE: A Autora (2018).
LEGENDA: Resultados da análise realizada com o banco de dados UNIPROTKB. Neste gráfico temos a representação da distribuição dos termos GO que mais ocorrem nos 21 *clusters* na categoria *Biological Process*

Outra característica observada nos resultados da categoria *Biological Process* é o fato de que os 21 clusters seguem o mesmo padrão com relação aos processos biológicos dos quais eles participam.

6.4 Mineração de Texto

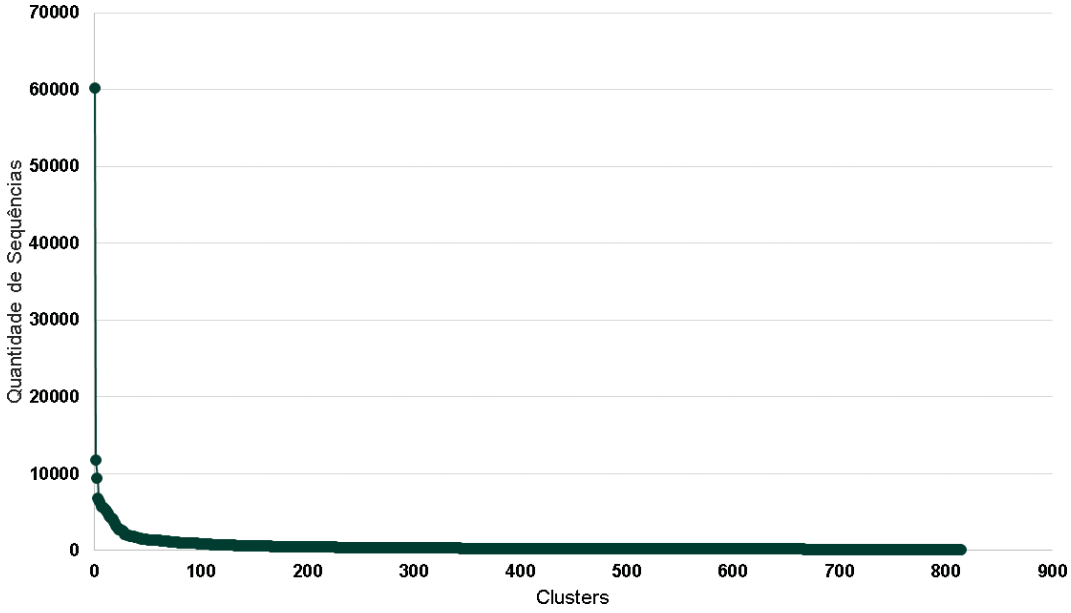
6.4.1 Análise dos clusters gerados na aplicação da mineração de texto

A clusterização do novo arquivo FASTA do cluster n° 1 criado após a transformação dos cabeçalhos para o formato de aminoácidos utilizando o algoritmo RAFTS^{3}G (NICHIO, 2016; NICHIO, et al., 2018) com valor de corte de 90% de identidade gerou 19.963 clusters, sendo que destes 10.658 apresentaram mais de 2 sequências. A clusterização demorou cerca de 3 horas para ser concluída.

Analisando a distribuição das 584.444 sequências nos 10.658 clusters com mais de 2 sequências (FIGURA 25) observou-se que:

I. A média de sequências por cluster é 8077;
II. O maior cluster é formado somente por proteínas hipotéticas, o que reforça a importância de analisar cuidadosamente os maiores clusters obtidos na clusterização de um banco de dados já que estes 10% de sequências compartilham 90% de identidade entre si e 50% de identidade com as demais as quais elas foram agrupadas na primeira clusterização.

FIGURA 25 - DISTRIBUIÇÃO DOS CLUSTERS COM MAIS DE DUAS SEQUÊNCIAS



FONTE: A Autora (2018).

6.4.2 Denominação dos Clusters Com Base na Mineração de Texto

Após a obtenção do arquivo contendo a estrutura dos *clusters* foi dado um nome padrão aos *clusters* com base no conteúdo textual dos Cabeçalhos. Os nomes padrão dados aos vinte maiores *clusters* foram listados na TABELA 6. O nome dado aos demais clusters não foi adicionado a TABELA 6 devido a inviabilidade de adicionar uma tabela de mais de dez mil linhas a esta dissertação. Porém, esta tabela pode ser acessada em: https://github.com/grazLet/DadosDisserta-o.

TABELA 6 - NOMES PADRÃO DOS 20 MAIORES CLUSTERS ENCONTRADOS NO FASTA CRIADO A PARTIR DA MINERAÇÃO DE TEXTO

Cluster ID	Nome padrão	N° de Membros
1	Hipothetical protein	60190
2	Transposase	11695
3	40S ribosomal protein S17	9386
4	Membrane protein	6742
5	Glucose-1-phosphate thymidylyltransferase	6494
6	Elongation factor 1-alpha, partial	6320
7	Vif protein	5651
8	ATP-dependent Clp protease proteolytic subunit	5618
9	Transposase, partial	5520
10	Glycerol kinase	5432
11	DNA-binding response regulator	5260
12	Tetr-family transcriptional regulator	5127
13	Transcriptional regulator	4884
14	Trna uridine 5-carboxymethylaminomethyl modification protein	4729
15	Endonuclease III	4410
16	Translation elongation factor 1-alpha, partial	4141
17	Elongation factor TU	4117
18	Peptide ABC transporter ATP-binding protein	4026
19	Cystathionine gamma-synthase	3601
20	Phosphoglucosamine mutase	3305

FONTE: A Autora (2018).

Analisando os nomes padrão dados aos *clusters*, verificamos que os nomes estão de acordo com os resultados obtidos tanto na ferramenta *PfamScan* quanto nas anotações funcionais do Gene Ontology, o que sugere a eficiência desta metodologia e da mineração de dados como auxiliar na interpretação de resultados obtidos tanto na clusterização quanto em análises de grandes conjuntos de sequências biológicas oriundos das mais diferentes fontes.

Comparando os resultados obtidos nas análises funcionais a partir dos dados do Gene Ontology, acreditamos que a grande quantidade de sequências classificadas no termo raiz das ontologias pode ser um reflexo da grande quantidade de sequências hipotéticas presentes no cluster n° 1.

6.5 O Pipeline

O pipeline de análise desenvolvido pode ser dividido em duas partes principais e um passo adicional:

I. Transferência de anotações do Gene Ontology com base em homologia de sequências;
II. Inferência de Homologia;
III. Nomeação padrão de grandes grupos de proteínas;

Para simplificar, o pipeline descrito aqui destina-se a usuários de diferentes ferramentas de clusterização e por isso, precisa apenas que o arquivo de entrada seja do tipo FASTA. Os dados utilizados neste trabalho são oriundos do banco de dados NR, porém o pipeline funciona para quaisquer espécies, bancos de dados e quantidade de sequências.

Caso os dados de entrada sejam muito grandes, como os utilizados no estudo de caso biológico descrito acima, é importante saber particionar sua análise de acordo com a quantidade de sequências que serão analisadas em cada etapa. A ferramenta *PfamScan* (LI, et al., 2015; MISTRY; BATEMAN; FINN, 2007) utiliza o HMMER3 (SEAN; WEELHER, 2018) e por isso, dependendo da quantidade de sequências do FASTA esta etapa pode demorar de dias a semanas, a não ser que o usuário possua uma *workstation* com múltiplos cores.

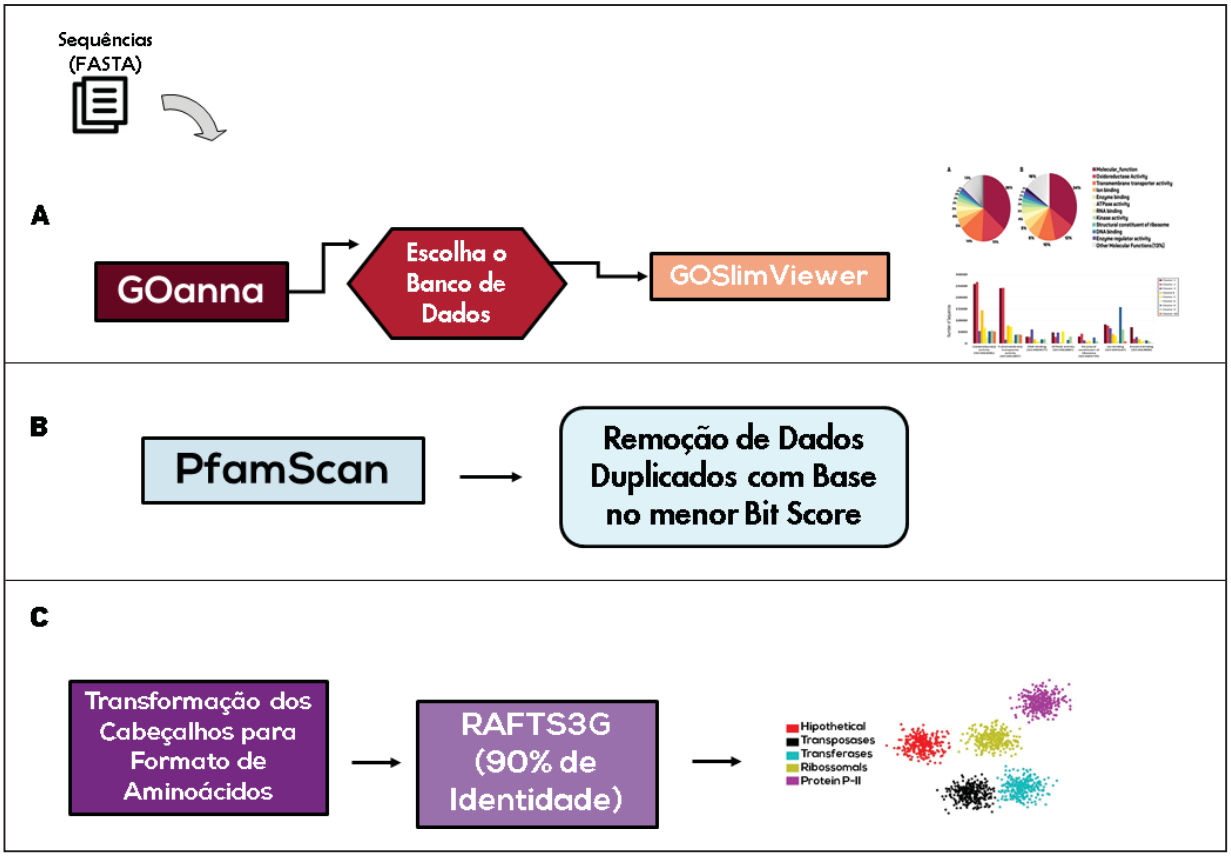
Assim, na primeira etapa do pipeline é realizada a transferência de anotações do Gene Ontology para as sequências do arquivo FASTA escolhido pelo usuário por meio da ferramenta GOanna. No site web da ferramenta o usuário pode escolher entre 13 bancos de dados, dentre os quais estão bancos específicos, como por exemplo, de fungos, plantas, pássaros, e também bancos que abrangem todas as espécies como o UniProtKB. A análise das sequências pode levar de minutos a algumas horas e então dentre os resultados obtidos, o arquivo “GOSummary” é a entrada para o próximo passo: sumarizar os resultados obtidos para então interpretá-los. A ferramenta GOSlimViewer utiliza o arquivo “GOSummary” como entrada e devolve ao usuário um

resumo utilizável em diversos aplicativos como Notepad++ e Excel.

O próximo passo do pipeline (FIGURA 2 B) é a inferência de homologia por meio da ferramenta PfamScan, que permite ao usuário descobrir informações biológico relevantes acerca do seu conjunto de dados como um todo, como por exemplo, inferir quais grupos de homólogos estão presentes no seu conjunto de dados, detectar indícios de homologia remota que podem guiar a interpretação dos dados e encontrar similaridades estruturais entre as proteínas.

A denominação dos clusters com base no conteúdo textual dos cabeçalhos é uma etapa adicional para facilitar a interpretação dos resultados. As funções necessárias para realização desta etapa podem ser encontradas em: https://github.com/grazLet/Give-a-cluster-name.

FIGURA 26 - FLUXOGRAMA DETALHANDO O FUNCIONAMENTO DO PIPELINE CRIADO NESTE TRABALHO



7 CONCLUSÕES E PERSPECTIVAS FUTURAS

Neste trabalho nós apresentamos a clusterização do banco de dados *Non-redundant Data Sequences of Proteins* (NR) utilizando o algoritmo RAFTS^{3}G e a análise dos 21 maiores clusters gerados utilizando técnicas de mineração de dados.

Analisando alguns aspectos da etapa de clusterização podemos concluir que o agrupamento realizado pelo algoritmo RAFTS^{3}G foi eficiente na resolução de problemas como armazenamento e busca contra o banco de dados, além de facilitar a curadoria dos dados.

As análises realizadas utilizando o conjunto de dados do cluster n° 1, permitiram a criação de um pipeline para reprocessamento de grandes grupos de proteínas. Este pipeline aplica diferentes técnicas de mineração de dados para encontrar novos padrões e interpretar os resultados brutos obtidos na clusterização. A princípio é possível afirmar que este pipeline é reprodutível pois o mesmo foi aplicado aos demais 20 maiores clusters do banco, dentre outros conjuntos de dados que não foram descritos neste trabalho.

Observando criteriosamente os resultados obtidos na análise que foi realizada para caracterizar os *clusters* com base nas anotações do *Gene Ontology* e do Pfam, é possível afirmar que os maiores *clusters* são compostos por proteínas e Famílias que exercem funções celulares essenciais encontradas em uma grande variedade de espécies, além de possuírem membros nos três Domínios da Vida.

Os resultados obtidos neste trabalho também trazem recomendações práticas aos usuários para usos mais eficazes dos resultados das ferramentas de clusterização de bancos de sequências biológicas.

Como perspectivas futuras pretendemos utilizar a metodologia criada para denominar os *clusters* na criação de um banco de dados online de sequências clusterizadas e integrar o pipeline ao banco para facilitar a interpretação dos resultados pelos usuários finais da informação. Também pretendemos automatizar todas as etapas do pipeline.

REFERÊNCIAS

ACEVEDO-ROCHA, C. G. et al. From essential to persistent genes: a functional approach to constructing synthetic life. **Trends in Genetics**, v. 29, n. 5. maio. 2013.

AGATONOVIC-KUSTRIN, S.; BERESFORD, S. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. **Journal of Pharmaceutical and Biomedical Analysis**, v. 22, n. 5, p. 717-27, jun. 2000.

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

ALQURAISHI, M.; TANG, S; XIA, X. An affinity-structure database of helix-turn-helix: DNA complexes with a universal coordinate system. **BMC Bioinformatics**, v. 16, n. 390. nov. 2015. **Journal of Biological Chemistry**, v. 284, p. 13519-13532. maio. 2009.

ANDERSSON, F. I. Structure and Function of a Novel Type of ATP-dependent Clp Protease.

ANDREEVA, A. et al. Data growth and its impact on the SCOP database: new developments. **Nucleic Acids Research**, v. 36, p. 419-25. nov. 2007.

ARAÚJO, G. F. **Codificação e Clustering de Proteínas**. 46 f. Trabalho de Graduação (Bacharelado em Ciência da Computação) – Departamento de Ciência da Computação, Universidade Federal de Lavras, Lavras, 2007.

ARAVIND, L.; KOONIN, E. V. DNA polymerase β-like nucleotidyltransferase superfamily: Identification of three new families, classification and evolutionary history. **Nucleic Acids Research**, v. 27, n. 7, p. 1609 – 1618. abr. 1999.

ARAVIND, L. et al. The many faces of the helix-turn-helix domain: Transcription regulation and beyond. **FEMS Microbiology Review**, v. 29, n. 2, p. 231-262. abr. 2005.

ARBELAITZ, O. et al. An extensive comparative study of cluster validity indices. **Pattern Recognition**, v. 46, p. 243-256. ago. 2012.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nature Genetics**, v. 25, n.1, p. 25-9, mai. 2000.

BALAKRISHNAN, R. et al. A guide to best practices for Gene Ontology (GO) manual annotation. **Database**, v. 2013. Jun. 2013.

BARKER, W. C. The Protein Information Resource (PIR). **Nucleic Acids Res**, v. 28, n.1, p. 41-44. jan. 2000.

BASHEER, I. A.; HAJMEER, M. Artificial neural networks: fundamentals, computing,

design, and application. **Journal of Microbiological Method**, v. 43, n. 1, p. 3-31, dez. 2000.

BATRA, V. K. et al. Amino acid substitution in the active site of DNA polymerase β explains the energy barrier of the nucleotidyl transfer reaction. **J Am Chem Soc**, v. 135, n. 21, p. 8078-8088. maio. 2013.

BEARD, W. A.; WILSON, A. H. Structure and Mechanism of DNA Polymerase β. **Biochemistry**, v. 53, n. 17, p. 2768-2780. abr. 2014.

BENSON, D. A. et al. GenBank. **Nucleic Acids Res**, v. 27, n.1, p. 12-7. jan. 1999.

BENSON, D. A. et al. GenBank. **Nucleic Acids Res,** v. 41, p. 36-42. nov. 2012.

BERMAN, H. M. et al. The Protein Data Bank. **Nucleic Acids Res**, v. 28, p. 235-242. jan. 2000.

BERNSTEIN, Y. Detection and Management of Redundancy for Information Retrieval. 210 f. Tese (Doutorado em Filosofia) - School of Computer Science and Information Technology, RMIT University, Melbourne, 2006.

BODEN, M. Alignment-free sequence comparison with spaced k-mers. **Open Access Ser. Inform**, v. 34, p. 24-34. jan. 2013.

BODENREIDER, O.; AUBRY, M.; BURGUN, A. Non-Lexical Approaches To Identifying Associative Relations In The Gene Ontology. **Pac Symp Biocomput**, p. 91-102. 2005.

BOURRET, R. B. Receiver domain structure and function in response regulator proteins. **Curr Opin Microbiol**, v. 13, n. 2, p. 142-149. mar. 2010.

BRENNAN, R.G.; MATTHEWS, B. W. The Helix-Turn-Helix DNA Binding Motif. **The Journal of Biological Chemistry**, v. 264, n. 4, p. 1903-6. fev. 1989.

BRODSKY, B.; PERSIKOV, A. V. Molecular Structure of the Collagen Triple Helix. **Advances in Protein Chemistry**, v. 70, p. 301-339. 2005.

BROWN, W. M. et al. Artificial neural networks: a new method for mineral prospectivity mapping. Australian Journal of Earth Sciences, v. 47, n. 4, p. 757-770, ago. 2000.

BUCHFINK, B.; XIE, C.; HUSON, D. H. Fast and Sensitive Protein Alignment using DIAMOND. **Nature Methods,** v. 12, p. 59-60. 2015.

BULL, S. C.; MULDOON. M. R.; DOIG, A. J. Maximising the Size of Non-Redundant Protein Datasets Using Graph Theory. **PLOS One**, v. 8, n. 2. fev. 2013.

BUCK, C. A. Immunoglobulin superfamily: structure, function and relationship to other receptor molecules. **Semin Cell Biol**, v. 3, n. 3, p. 179-88. jun. 1992.

CAMACHO, C. et al. BLAST+: architecture and applications. **BMC Bioinformatics**, v. 10, n. 421. dez. 2009.
CAETANO-ANOLLES, G.; CAETANO-ANOLLES, D. An evolutionary structured universe of protein architecture. **Genome Res**, v. 13, n. 7, p. 1563 – 1571. jul. 2003.

CAETANO-ANOLLES, G.; CAETANO-ANOLLES, D. Universal sharing patterns in proteomes and evolution of protein fold architecture and life. **Mol. Evol**, v. 60, n. 4, p. 484-498. abr. 2005.

CASTANHEIRA, L. G. **Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões**. 95 f. Dissertação (Mestrado em Engenharia Elétrica) – Departamento de Engenharia Elétrica, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

CENTER FOR BIOINFORMATICS & COMPUTATIONAL BIOLOGY. Spring 2010. Disponível em: Acesso em 15 ago. 2018.

CERRITELLI, S. M.; CROUCH, R. J. Ribonuclease H: the enzymes in Eukaryotes. **FEBS J**, v. 276, n. 6, p. 1494-1505. mar. 2008.

CHAWLA, S.; SHARMA, R. Application of Data Mining in Bioinformatics. **International Journal of Engineering Science and Computing**, v. 6, n. 6, p. 7426-9. jun. 2016.

CHEN, Q. et al. Comparative Analysis of Sequence Clustering Methods for Deduplication of Biological Databases. **ACM Journal of Data and Information Quality**, v.9, n. 3. jan. 2018.

COHEN, K. B.; HUNTER, L. Getting Started in Text Mining. **PLoS Computational Biology**, v. 4, n.1, p. 1-3, jan. 2008.

DEAN, N.; RZHETSKY, A.; ALLIKMETS, R. The human ATP-binding cassette (ABC) transporter superfamily. **Genome Res**, v. 11, n. 7, p. 1156-66. jul. 2001.

DURBIN, R. et al. **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids**. Cambridge: University of Cambridge, 1998

EDDY, S. R. Profile hidden Markov models. **Bioinformatics**, v. 14, n. 9, p.755-763. out. 1998.

EDDY, S. R. Accelerated Profile HMM Searches. **PLoS Comput Biol**, v. 7, n.10. out. 2011.

EDGAR, R. C. Search and clustering orders of magnitude faster than BLAST. **Bioinformatics**, **v. 26, n.19, p. 2460–2461. out. 2010.**

EMMS, D. M.; KELLY, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. **Genome Biology**, v. 16, n.1. ago. 2015.

ERICKSON, H. P. Evolution of the cytoskeleton. **BioEssays**, v. 29, p. 668-667. jun. 2007.

FINN, R. D. et al. The Pfam protein families database: towards a more sustainable future. **Nucleic Acids Research**, v. 44, p. 279-285. jan. 2016.

FINN, R. D. et al. Pfam: clans, web tools and services. **Nucleic Acids Research,** v. 34, p. 247-251. jan. 2006.

FISLAGE, M.; WAUTERS, L.; VERSÉES, W. Invited review: MnmE, a GTPase that drives a complex tRNA modification reaction. **Biopolymers**, v. 105, n. 8, p. 568-79. ago. 2016.

FU, L. et al. CD-HIT: accelerated for clustering the next generation sequencing data. **Bioinformatics**, v. 28, n. 23, p. 3150-2. out. 2012.

GARDNER, M. W.; DORLING, S. **Artificial Neural Networks (The Multilayer Perceptron) – A Review of Applications in the Atmospheric Sciences.** Atmospheric Environment**, v. 32, n. 14/15, p. 2627-2636, ago. 1998.**

GRONT, D.; KOLINSKI, A. HCPM—program for hierarchical clustering of protein models. **Bioinformatics Applications Note**, v. 21, n. 14, p. 3179–3180, abr. 2005.

GANTEN, D. Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine. In: KWON, M. S.; CHO, S. Y.; PAIK, Y. K. **Protein Databases**. Nova Iorque: Springer, 2006. p. 1483-1487.

GASTEIGER, E. et al. ExPASy: The proteomics server for in-depth protein knowledge and analysis. **Nucleic Acids Res,** v. 31, n.13, p. 3784-8. jul. 2003.

GENE ONTOLOGY CONSORTIUM. The Gene Ontology (GO) database and informatics resource. **Nucleic Acids Research**, v. 32, n. 1, p. 258-262. jan. 2004.

GENE ONTOLOGY CONSORTIUM. Gene Ontology Consortium: going forward. **Nucleic Acids Research,** v. 43, p. 1049-1056. nov. 2014.

GHAHRAMANI, Z. An Introduction to Hidden Markov Models and Bayesian Networks. **International Journal of Pattern Recognition and Artificial Intelligence**, v. 15, n. 1, p. 9-42.

GO Consortium (2018) Guide to GO Evidence Codes. Disponível em:< http://www.geneontology.org/GO.evidence.shtml>. Acesso em: 10 mar. 2018.

GONZALEZ, G. H. et al. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. **Briefings in Bioinformatics**, v. 17, n. 1, p. 33-42, set. 2015.
GUNASEKARAN, M. et al. Machine Learning Based Big Data Processing Framework for Cancer Diagnosis Using Hidden MarkovModel and GM Clustering. **Wireless Pers Commun**, p. 1-18. nov. 2017.

GUNNING, P. W. et al. The evolution of compositionally and functionally distinct actin filaments. **Journal of Cell Science**, v. 128, n. 11, p. 2009-2019. jun. 2015.

HAFT, D. H. et al. TIGRFAMS and Genome Properties in 2013. **Nuclei Acids Research,** v. 41, p. 387-395. nov. 2012.

HAFT, D. R.; HAFT, D. H. A comprehensive software suite for protein family construction and functional site prediction. PLoS One, v.12, n. 2. fev. 2017.

HARTMANN, G. **Advances in Immunology.** Elsevier, 2017. Ebook. Disponível em: < https://www.sciencedirect.com/journal/advances-in-immunology/issues>. Acesso em: 20 out. 2018.

HERMAN, I. M. Actin Isoforms. **Curr Opin Cell Biol**, v. 5, n. 1, p. 48-55. fev. 1993.

HIPP, R. et al. **SQLite (Version 3.8.10.2)**. SQLite Development Team. Disponível em <https://www.sqlite.org/download.html>. Acesso em: 15 ago. 2017.

HOLM, L; SANDER, C. Removing near-neighbour redundancy from large protein sequence collections. **Bioinformatics**, v. 14, n. 5, p. 423-9, jan. 1998.

HOLZINGER, A.; JURISICA, I. Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions. **Lecture Notes in Computer Science**, v. 8401, p. 1-18. 2014.

HOLZINGER, A.; DEHMER, M.; JURISICA, I. Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. **BMC Bioinformatics**, v. 15, n.6, p. 1-9. 2014.

HUAIYU, M. et al. Large-scale gene function analysis with the PANTHER classification system. **Nature Protocols,** v. 8, n.8, p. 1551-66. ago. 2013.

HUAIYU, M. et al. PANTHER version 10: expanded protein families and functions and analysis tools. **Nucleic Acids Research,** v. 44, p.336-342. nov. 2015.

HUAIYU, M. et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. **Nucleic Acids Research**, v. 45, p. 183-189. nov. 2016.

HUSTEN, E. J.; EIPPER, B. A.

KABSCH, W.; VANDERKERCKHOVE, J. Structure and function of actin. **Annu Rev Biophys Biomol Struct**, v. 21, p. 49-76. 1992.
KAWAMURA, Y. et al. Systematic Analyses of P-Loop Containing Nucleotide Triphosphate Hydrolase Superfamily Based on Sequence, Structure and Function. **Genome Informatics**, v. 14, p. 581-2. 2003.

KELIL, A. et al. CLUSS: Clustering of protein sequences based on a new similarity measure. **BMC Bioinformatics**, v. 8, n. 286, p. 1471- 2105. aug. 2007.

KIM, E. et al. Extracellular Domain of V-Set and Immunoglobulin Domain Containing 1 (VSIG1) Interacts with Sertoli Cell Membrane Protein, while Its PDZ-Binding Motif Forms a Complex with ZO-1. **Mol Cells**, v. 30, n. 5, p. 443-448. nov. 2010.

KIM, C. S. et al. K-mer clustering algorithm using a MapReduce framework: application to the parallelization of the Inchworm module of Trinity. **BMC Bioinformatics**, v. 18, n. 1, p. 467. nov. 2017.

KRAB, I. M.; PARMEGGIANI, A. Mechanisms of EF-Tu, a pioneer GTPase. **Progress in Nucleic Acid Research and Molecular Biology**, v. 71, p. 513-551. 2002.

KROGH, A. et al. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. **Journal of Molecular Biology,** v. 235, n. 5, p. 1501-1531. fev. 1994.

LEIMEISTER, C. A. et al. Fast alignment-free sequence comparison using spaced-word frequencies. **Bioinformatics**, v. 30, p. 1991-1999. jul. 2014.

LETKO, M. et al. Identification of the HIV-1 Vif and human APOBEC3G protein interface. **Cell Rep**, v. 13, n. 9, p. 1789-1799. dez. 2015.

LETUNIC, I.; DOERKS, T.; BORK, P. SMART: recent updates, new developments and status in 2015. **Nucleic Acids Research,** v. 43, p. 257-260. oct. 2014.

LETUNIC, I.; BORK, P. 20 years of the SMART protein domain annotation resource. **Nucleic Acids Research**, v. 46, p. 493-6. oct. 2017.

LI, R.; LI, X. Q.; WANG, G. Improved and Novel Cluster Analysis for Bioinformatics, Computational Biology and All Other Data. **Int'l Conf on Bioinformatics & Computational Biology. BIOCOMP'15**. Disponível em: < http://worldcomp-proceedings.com/proc/p2015/BIC1404.pdf>. Acesso em: 22 setembro 2017.

LI, W.; JAROSZEWSKI, L.; GODZIK, A. Clustering of highly homologous sequences to reduce the size of large protein databases. **Bioinformatics Applications Note**, v. 17, n. 3, p. 282-3, out. 2000.

LI, W.; JAROSZEWSKI, L.; GODZIK, A. Sequence clustering strategies improve remote homology recognitions while reducing search times. **Protein Engineering**, v. 15, n. 8, p.

643-9. 2000.

LI, W.; JAROSZEWSKI, L.; GODZIK, A. Tolerating some redundancy significantly speeds up clustering of large protein databases. **Bioinformatics**, v. 18, n. 1, p. 77-82. 2002.

LI, W. et al. The EMBL-EBI bioinformatics web and programmatic tools framework. **Nucleic Acids Res**, v. 1, n. 43, p. 580-4. abr. 2015.

LOEWENSTEIN, Y. et al. Protein function annotation by homology-based inference. **Genome Biology**, v. 10, n. 2. fev. 2009.

MA, B.-G. et al. Characters of very ancient proteins. **Biochemical and Biophysical Research Communications**, v. 366, n. 3, p. 607-611. fev. 2008.

MACCUISH, J.D.; MACCUISH, N.E. **Clustering in Bioinformatics and Drug Discovery**. Nova Iorque: CRC Press, 2011.

MANNA, S. An overview of pentatricopeptide repeat proteins and their applications. **Biochimie**, v. 113, p. 93-9. jun. 2015.

McCARTHY, F. M. AgBase: a functional genomics resource for agriculture. **BMC Genomics**, v. 7, n. 229. set. 2006.

McEVER, R. P.; LUSCINSKAS, F. W. **Hematology:** Basic Principles and Pratic. Filadélfia: Elsevier, 2018.

MASOOD, M. A.; KHAN, M. N. A. Clustering Techniques in Bioinformatics**. I.J. Modern Education and Computer Science**, v. 1, p. 38-46, jan. 2015.

MAURIZI, M. R. et al. Sequence and structure of Clp P, the proteolytic component of the ATP-dependent Clp protease of Escherichia coli. **J Biol Chem**, v. 265, n. 21, p. 12536-45. jul. 1990.

MAYNE, R.; BREWTON, R. G. New Members of the collagen superfamily. **Current Opinion in Cell Biology**, v. 5, n. 5, p. 883-890. out. 1993.

MAZANDU, G. K. et al. A-DaGO-Fun: an adaptable Gene Ontology semantic similarity-based functional analysis tool. **Bioinformatics**, v. 32, n. 3, p. 477-9. out. 2015.

MAZANDU, G. K.; CHIMUSA, E. R.; MULDER, N. J. Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge Discovery. **Briefings in Bioinformatics,** v. 18, n. 5, p. 886-901. jul. 2016.

MAZANDU, G. K.; MULDER, N. J. DaGO-Fun: tool for Gene Ontology-based functional analysis using term information content measures. **BMC Bioinformatics**, v. 14. set. 2013.

MISTRY, J.; BATEMAN, A.; FINN, R. D. Predicting active site residue annotations in the Pfam database. **BMC Bioinformatics**, v. 8, n. 298. ago. 2007.
MURZIN, A. G. et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. **J. Mol. Biol**, v. 247, n. 4, p. 536-40. abr. 1995.

NICHIO, B. T. L. **Consolidação e validação da ferramenta Rapid Alignment Free Tool for Sequences Similarity Search to Groups (RAFTS3groups) – Um software rápido de clusterização para Big Data e buscador consistente de proteínas ortólogas**. Dissertação (Mestrado em Bioinformática) –Universidade Federal do Paraná, Curitiba, 2016.

NICHIO, B. T. L. RAFTS³G – An efficient and versatile clustering software to analyses in large protein datasets. **bioRxiv 407437**. set. 2018.

NEEDLEMAN, S.B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **J Mol Biol**, v. 48, n. 3, p. 443-453. mar. 1970.

OLDEHINKEL, E. B.; DOEVEN, M. K.; POOLMAN, B. ABC transporter architecture and regulatory roles of accessory domains. **FEBS Letters**, v. 580, n. 4, p. 1023 – 1035. fev. 2016.

PABO, C. O.; SAUER, R. T. Protein-DNA Recognition. **Review of Biochemistry**, v. 53, n. 1, p. 293-321. jan. 1984.

PAN, T. et al. Kmerind: A Flexible Parallel Library for K-mer Indexing of Biological Sequences on Distributed Memory Systems. **IEEE/ACM Trans Comput Biol Bioinform,** out. 2017.

PEARSON, W. R. An introduction to sequence similarity (“homology”) searching.
Current Protocols in Bioinformatics, n. SUPPL.42, p. 1–8, 2013.

PUNTA, M.; MISTRY, J. **Homology-Based Annotation of Large Protein Datasets**, In: **Data Mining Techniques for the Life Sciences**. **Methods in Molecular Biology**. Nova Iorque: Humana Press, 2016.

REUTER, J. A.; SPACEK, D.; SNYDER, M. P. High-Throughput Sequencing Technologies. **Mol Cell,** v. 58, n.4, p. 586-597. maio. 2015.

RICE, P.; LONGDEN, I.; BLEASBY, A. EMBOSS: the European Molecular Biology Open Software Suite. **Trends Genet,** v. 16, n.6, p. 276-7. jun. 2000.

RODRIGUES, T. S. et al. Clustering and artificial neural networks: Classification of variable lengths of Helminth antigens in set of domains. **Genetics and Molecular Biology**, v. 27, n. 4, p. 673-8, ago. 2004.

ROGERS, M. F. BEN-HUR, A. The use of gene ontology evidence codes in preventing classifier assessment bias. **Bioinformatics**, v. 25, n. 9, p. 1173-1177. maio. 2009.
ROSE, K. M. et al. The viral infectivity factor (Vif) of HIV-1 unveiled. **Trends in Molecular Medicine**, v. 10, n. 6, p. 291-297. jun. 2004.

ROSSUM, G. V.; DRAKE, F. L. Python 3 Reference Manual. Paramount: California, 2009.

RUSSEL, S.; NORVIG, P. **Inteligência Artificial.** Rio de Janeiro: Elsevier, 2013.

SARAVANAN, N.; DEVI, T. A Survey On Biological Databases And Applications Of Datamining. **Australian Journal of Basic and Applied Sciences**, v. 6, n. 13, p. 175-180. 2012.

SAUER, R. T. et al. Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. **Nature**, v. 298, n. 5873, p. 447-451. jul. 1982.

SHA, J. et al. Molecular Characterization of a Glucose-Inhibited Division Gene, gidA, That Regulates Cytotoxic Enterotoxin of Aeromonas hydrophila. **Infect Immun**, v. 72, n.2, p. 1084-1095. fev. 2004.

SLACK, F. J.; RUVKUN, G. A novel repeat domain that is often associated with RING finger and B-box motifs. **Trends Biochem Sci**, v. 23, n. 12, p. 474-475. dez. 1998.

SERGIEV, P. V.; BOGDANOV, A. A.; DONTSOVA, O. A. How can elongation factors EF-G and EF-Tu discriminate the functional state of the ribosome using the same binding site? **FEBS Letters**, v. 579, n. 25, p. 5439-5442. set. 2005.

SIKIC. K.; CARUGO, O. Protein sequence redundancy reduction: comparison of various method. **Bioinformation,** v. 5, n. 6, p. 234-239. nov. 2010.

SMITH, T. F.; WATERMAN, M.S. Identification of common molecular subsequences. **Journal of Molecular Biology**, v. 147, n. 1, p. 195-197. mar. 1981.

TRAIN. Primary and secondary databases. Disponível em: < https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified-2018/primary-and-secondary-databases>. Acesso em: 28 ago. 2018.

THE UNIPROT CONSORTIUM. UniProt: the universal protein knowledgebase. **Nucleic Acids Research**, v. 46, n. 5, p. 2699. mar. 2018. abr. 2001.

TODD, A. E.; ORENGO, C. A.; THORNTON, J. M. Evolution of Function in Protein Superfamilies, from a Structural Perspective. **J. Mol. Biol**, v. 307, p. 1113-1143.

ULLMAN, J. D.; WIDOM, J. **A First Course in Database Systems**. Nova Jersey: Upper Saddle River, 1997.

VASILIOU, V.; VALISIOU, K.; NEBERT, D. Human ATP-binding cassette (ABC) transporter Family. **Hum Genomics**, v. 3, n. 3, p. 281-290. abr. 2009.

VIALLE, R. A. et al. RAFTS3: Rapid Alignment-Free Tool for Sequence Similarity Search. **bioRxiv**. mai. 2016.

XU, D. Protein Databases on the Internet. **Curr Protoc Mol Biol**, p. 19-4. jan. 2014.

WAGNER, S. C. **Biological Nitrogen Fixation.** Nature Education Knowledge**,** v. 3, n.10. 2011.

WALKER, J. E. et al. Distantly related sequence in the a and b subunits of ATP synthase, myosin, kinase and other ATP requiring enzymes and a common nucleotide binding fold. **EMBO J**, v. 1, n.8, p. 945- 951. 1982.

WANG, G. DUNBRACK, R. L. Jr. PISCES: recent improvements to a PDB sequence culling server. **Nucleic Acids Research,** v. 33. jul. 2005.

WEBB, E. C. **Enzyme Nomenclature**: 1992 Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. Nova Iorque: Academic Press, 1992.

WHITE, D. J. et al. GidA is an FAD-binding protein involved in development of Myxococcus xanthus. **Molecular Microbiology**, v. 42, n. 2, p. 503-517. jul. 2008.

WOSZEZENKI, C. R.; GONÇALVES, A. L. Biomedical text mining: a bibliometrics review. **Perspectivas em Ciência da Informação**, v. 18, n. 3, p. 24-44, jul./set. 2013.

ZHAO, W.; ZOU, W.; CHEN, J. J. Topic modeling for cluster analysis of large biological and medical datasets. **BMC Bioinformatics**, v. 15, n. 11, p. 1-11, mar. 2014.

ZOU, D. et al. Biological Databases for Human Research. **Genomics Proteomics Bioinformatics**, v. 13, p. 55-63. fev. 2015.

APÊNDICE

APÊNDICE 1 – Função Cataunicos

```
function [idseq,CC, maiorvalor] = cataunicos(n,tabelaCell,tabelaNumerica, index)
idseq=unique(tabelaCell(index==n,1));
CC=unique(tabelaCell(index==n,2));
maiorvalor=num2cell(max(tabelaNumerica(index==n,1)));
end
```

APÊNDICE 2 – Função *dna2list*

```
function mret = w_tata(sdna, quant)
n = length(sdna);
inds = [];
for i=1:quant
    inds = [inds (i:(n-quant+i))'];
end
mret = sdna(inds);
if quant==1
    mret = mret';
end
```

APÊNDICE 3 – Função *getheadersfeats.m*

```
% Get features
%NS = ResumoClus.NS;
S = cellfun(@(x) char(last(getstrentre(x,'|',' [',0))),NS(1:1249),'UniformOutput',false);
S = cleantext(trimall(S)); %data in
%S = cellfun(@(x) x(22:first(strfind(upper(x),'OS=')-1)),S,'UniformOutput',false );
S =
termoff(S,{'protein','subunit','family','MULTISPECIES','PUTATIVE','HYPOTHETICAL','PA
RTIAL','PREDICTED'});%load Rtx.mat
S = trimall(S);
%L = 27^4;
%Rtx = orthbase(L,729);
tic; Wtx = text2mat(S,Rtx); toc
%Got
ids = sorti(rand(length(Wtx(:,1)),1));
wr= Wtx(ids,:);
S4 = S(ids(1:8000));
[cl0 b net0 c d e f D x y Zs] = HFclus(wr(1:8000,:),'Loops',300); %
%
%load cltr
n = length(Wtx(:,1));
cl1 = double(ismember(cltr,[7 9]));
cl2 = 2*double(ismember(cltr,[2 3 4 5 6]));
cl3 = 3*double(~(cl1+cl2));
clnet = cl1+cl2+cl3;
%
wall = [wr(1:8000,:) clnet];
wtr = wall(1:7000,:);
wts = wall(7001:8000,:);
r = mlp_tr(wtr,[7 5 ]);
q = mlp_ts(wtr,r)
```

APENDICE 4 – Script utilizado na atribuição de nome padrão aos clusters

```
function [ cluster ] = rafts_clusterFROMxfas( xfas, rftgroups, N)
%clusterFROMxfas obtem todas as sequências presentes em xfas que compoem o cluster N
    %rftgroups deve ser a struct do resultantes do raft3groups de xfas
    cluster=xfas(rftgroups.igrp==N);
    %cluster = readfastadirectp2(dbstruct.rdfs,find(rftgroups.contall(rftgroups.contall(:,1)==N,1)==rftgroups.igrp),dbstruct.rdid);
end

T = struct2cell(cluster)';
cluster_headers = T(:,1);
cluster_name = text2mat(cluster_headers);
WT = cluster_name;
[a b] = distminWc(mean(WT),WT)
a
cluster11(13).Header
```

APENDICE 5 – Função text2mat

```
function [Z R] = text2mat2(TX, varargin)
% Versão atualizada de text2vetc (05-07-2018)
% Gera o vetor baseados em spaced words representando o string em TX
% O formato da spaced word: x0xxx
% Projeta na base R
L = 531441; %27^4
n = length(TX);
if isempty(varargin)
   R = orthbase(L,729);
else
   R = varargin{1};
end
nr = length(R(1,:));
Z = zeros(n,nr);
%
istep = 10;
for ii=1:istep:n
   iend = (ii+istep-1);
   iend = min(n,iend);
   disp(num2str([ii iend]));
   Z(ii:iend,:) = cell2mat(cellfun(@(x) (tx2mat(x)),TX(ii:iend),'UniformOutput',false))*R;
end
end

function vout = tx2mat(xseq)
% xseq - string
% varargin - sampling length. Eg. 1 monopeptide, 2 dipeptide...
m = 531441; %27^4
vout = zeros(1,m);
xseq = upper(trimall(bs2b(onlyletters(xseq,1))));
xseq(xseq==' ') = 64;
xseq = xseq-64;
L5 = dna2list(xseq,5);
l = length(L5(:,1));
L4 = [L5(:,1) L5(:,3:end)];
pots = repmat([1 27 729 19683],l,1);
ids = sum((L4.*pots)')+1;
vout(ids) = 1;
end
```

APÊNDICE 6 – Função distminWC

```
function [cl D] = distminWc(w,c)
[nw m] = size(w);
nc = length(c(:,1));
p1 = vet2mat(mat2vet(repmat(w',nc,1)'),m);
p2 = repmat(c,nw,1);
D = vet2mat(normavect(p1-p2),nc);
cl = mini(D')';
```

APÊNDICE 7 - DESCRIÇÃO COMPLEMENTAR DOS DADOS DOS CLÃS DO PFAM

ID	Número de Sequências	ID	Número de Sequências	ID	Número de Sequências	ID	Número de Sequências
CL0011	33553	**CL0151**	2541	**CL0027**	1418	**CL0196**	735
CL0123	28338	**CL0591**	2540	**CL0067**	1409	**CL0260**	725
CL0219	23063	**CL0107**	2419	**CL0109**	1389	**CL0094**	683
CL0023	20269	**CL0106**	2387	**CL0198**	1384	**CL0352**	679
CL0063	19941	**CL0136**	2198	**CL0220**	1286	**CL0533**	640
CL0036	8115	**CL0184**	2140	**CL0255**	1272	**CL0304**	609
CL0108	7605	**CL0089**	2134	**CL0113**	1257	**CL0153**	604
CL0548	7179	**CL0177**	2101	**CL0072**	1201	**CL0541**	599
CL0231	6361	**CL0029**	2042	**CL0341**	1125	**CL0261**	590
CL0254	6299	**CL0022**	2038	**CL0279**	1114	**CL0051**	583
CL0016	5051	**CL0257**	1912	**CL0531**	1103	**CL0615**	577
CL0288	4448	**CL0364**	1859	**CL0329**	1080	**CL0010**	570
CL0021	4341	**CL0014**	1837	**CL0020**	1078	**CL0387**	553
CL0172	4262	**CL0343**	1799	**CL0344**	1051	**CL0040**	552
CL0492	4237	**CL0378**	1721	**CL0033**	1023	**CL0269**	545
CL0062	4130	**CL0270**	1713	**CL0190**	997	**CL0502**	527
CL0105	3546	**CL0167**	1604	**CL0292**	990	**CL0382**	509
CL0465	3528	**CL0649**	1602	**CL0199**	915	**CL0266**	509
CL0031	3420	**CL0181**	1549	**CL0034**	897	**CL0233**	497
CL0179	3318	**CL0104**	1524	**CL0323**	858	**CL0039**	494
CL0046	3010	**CL0070**	1520	**CL0337**	833	**CL0058**	477
CL0061	2984	**CL0314**	1494	**CL0149**	831	**CL0124**	472
CL0404	2671	**CL0035**	1469	**CL0174**	766	**CL0423**	452
CL0588	2643	**CL0296**	1452	**CL0237**	754	**CL0408**	444

APÊNDICE 7 - DESCRIÇÃO COMPLEMENTAR DOS DADOS DOS CLÃS DO PFAM

ID	Número de Sequências	ID	Número de Sequências	ID	Número de Sequências	ID	Número de Sequências
CL0050	438	**CL0006**	292	**CL0525**	179	**CL0536**	103
CL0137	429	**CL0336**	284	**CL0469**	173	**CL0297**	103
CL0366	425	**CL0572**	283	**CL0048**	170	**CL0159**	102
CL0041	425	**CL0110**	279	**CL0325**	168	**CL0287**	95
CL0192	423	**CL0057**	274	**CL0235**	159	**CL0232**	95
CL0334	421	**CL0251**	272	**CL0114**	159	**CL0262**	94
CL0481	418	**CL0142**	264	**CL0399**	158	**CL0310**	92
CL0487	417	**CL0441**	255	**CL0316**	154	**CL0204**	91
CL0280	391	**CL0350**	241	**CL0306**	153	**CL0182**	89
CL0122	387	**CL0494**	236	**CL0503**	147	**CL0530**	89
CL0049	382	**CL0132**	232	**CL0303**	144	**CL0347**	83
CL0222	377	**CL0028**	229	**CL0225**	134	**CL0397**	81
CL0396	373	**CL0431**	229	**CL0608**	132	**CL0328**	80
CL0015	363	**CL0505**	222	**CL0604**	127	**CL0236**	80
CL0436	360	**CL0171**	218	**CL0188**	126	**CL0064**	79
CL0221	359	**CL0025**	216	**CL0214**	119	**CL0158**	78
CL0098	356	**CL0534**	214	**CL0144**	117	**CL0127**	75
CL0032	344	**CL0349**	211	**CL0207**	114	**CL0600**	74
CL0248	341	**CL0193**	200	**CL0363**	110	**CL0131**	74
CL0486	338	**CL0206**	191	**CL0504**	110	**CL0018**	74
CL0247	299	**CL0186**	189	**CL0126**	110	**CL0201**	71
CL0145	299	**CL0268**	186	**CL0030**	109	**CL0053**	61
CL0381	297	**CL0042**	181	**CL0170**	107	**CL0346**	61
CL0116	295	**CL0081**	180	**CL0489**	106	**CL0044**	61

APÊNDICE 7 - DESCRIÇÃO COMPLEMENTAR DOS DADOS DOS CLÃS DO PFAM

ID	Número de Sequências	ID	Número de Sequências	ID	Número de Sequências
CL0209	60	**CL0228**	19	**CL0101**	4
CL0529	59	**CL0645**	17	**CL0148**	4
CL0340	55	**CL0318**	16	**CL0633**	3
CL0299	54	**CL0154**	14	**CL0045**	3
CL0654	50	**CL0311**	14	**CL0527**	3
CL0003	47	**CL0129**	13	**CL0083**	3
CL0490	41	**CL0259**	12	**CL0613**	3
CL0498	40	**CL0084**	12	**CL0118**	3
CL0320	40	**CL0252**	11	**CL0059**	3
CL0130	39	**CL0246**	11	**CL0472**	3
CL0511	38	**CL0004**	10	**CL0595**	2
CL0112	37	**CL0516**	9	**CL0508**	2
CL0523	36	**CL0327**	8	**CL0609**	2
CL0163	36	**CL0012**	8	**CL0348**	2
CL0300	33	**CL0357**	8	**CL0652**	2
CL0007	31	**CL0497**	7	**CL0075**	2
CL0475	27	**CL0406**	7	**CL0575**	2
CL0290	26	**CL0161**	7	**CL0517**	1
CL0556	25	**CL0418**	7	**CL0090**	1
CL0433	25	**CL0085**	6	**CL0092**	1
CL0080	24	**CL0074**	5	**CL0351**	1
CL0263	23	**CL0095**	5	**CL0298**	1
CL0073	20	**CL0388**	5	**CL0054**	1
CL0178	20	**CL0079**	5	**CL0449**	1
CL0466	19	**CL0229**	5		