UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática

Pâmela Marinho Rezende

AVALIAÇÃO DE ABORDAGENS HIERÁRQUICAS DE APRENDIZADO DE MÁQUINA APLICADAS A BANCOS DE DADOS BIOLÓGICOS

Belo Horizonte
2022

Pâmela Marinho Rezende

AVALIAÇÃO DE ABORDAGENS HIERÁRQUICAS DE APRENDIZADO DE MÁQUINA APLICADAS A BANCOS DE DADOS BIOLÓGICOS

Versão final

Tese apresentada ao Programa Interunidades de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Doutora em Bioinformática

Orientador: Ph.D. Douglas Eduardo Valente Pires
Co-orientador: Ph.D. Gabriel da Rocha Fernandes

Belo Horizonte
2022

043 Rezende, Pâmela Marinho.
Avaliação de abordagens hierárquicas de aprendizado de máquina aplicadas a bancos de dados biológicos [manuscrito] / Pâmela Marinho Rezende. – 2022.
172 f. : il. ; 29,5 cm.

Orientador: Ph.D. Douglas Eduardo Valente Pires. Co-orientador: Ph.D. Gabriel da Rocha Fernandes.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Aprendizado de Máquina. 3. Base de Dados. 4. Biologia. 5. Classificação automática. I. Pires, Douglas Eduardo Valente. II. Fernandes, Gabriel da Rocha. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DE DEFESA DE TESE

PÂMELA MARINHO REZENDE

Às dezessete horas do dia **25 de julho de 2022**, reuniu-se, através de videoconferência, a Comissão Examinadora de Tese indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho de Pâmela Marinho Rezende intitulado: **"AVALIAÇÃO DE ABORDAGENS HIERÁRQUICAS DE APRENDIZADO DE MÁQUINA APLICADOS A BANCOS DE DADOS BIOLÓGICOS"**, requisito para obtenção do grau de Doutora em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Douglas Eduardo Valente Pires**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa da candidata. Logo após, a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor(a)/ Pesquisador(a)	Instituição	Indicação
Dr. Douglas Eduardo Valente Pires	University of Melbourne	Aprovada
Dr. Gabriel da Rocha Fernandes	Instituto René Rachou, Fundação Oswaldo Cruz	Aprovada
Dra. Glaura da Conceição Franco	Universidade Federal de Minas Gerais	Aprovada
Dr. Laurence Rodrigues do Amaral	Universidade Federal de Uberlândia	Aprovada
Dra. Fabíola Souza Fernandes Pereira	Universidade Federal de Uberlândia	Aprovada

Pelas indicações, a candidata foi considerada: **Aprovada**

O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 25 de julho de 2022.

Documento assinado eletronicamente por **Glaura da Conceição Franco**, **Chefe de departamento**, em 25/07/2022, às 18:56, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gabriel da Rocha Fernandes**, **Usuário Externo**, em 25/07/2022, às 18:56, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Laurence Rodrigues do Amaral**, **Usuário Externo**, em 25/07/2022, às 18:58, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Douglas Eduardo Valente Pires**, **Usuário Externo**, em 25/07/2022, às 20:10, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fabíola Souza Fernandes Pereira**, **Usuária Externa**, em 26/07/2022, às 11:40, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1589765** e o código CRC **D78BD5B7**.

Referência: Processo nº 23072.240311/2022-85 SEI nº 1589765

Dedico este trabalho ao Danilo que foi e é meu suporte em toda minha carreira.

AGRADECIMENTOS

Aos meus pais, Regina e Sérgio que sempre me incentivaram nos meus estudos.
Ao meu esposo Danilo pelo apoio em toda a minha carreira acadêmica e profissional.
A minha irmã Marcela em todo o carinho e apreço, e também ao meu cunhado Daniel, Lavínia e Nicole por todo amor.
Ao PPG em Bioinformática da UFMG pela oportunidade em desenvolver o meu doutorado. A CAPES por fomentar essa pesquisa.
Ao meu orientador Douglas por todo o acompanhamento, direcionamento, escuta e empatia em todo o doutorado.
Ao meu co-orientador Gabriel pelo gatilho inicial da temática do trabalho e o apoio ao projeto.
Aos meus colegas da Fiocruz da Plataforma de Bioinformática, Joicy, João, Juliana, Amanda, Frã e Fausto pelos compartilhamentos de conhecimentos, comida e boas risadas. Em especial a Joicera que acompanhou e me incentivou em toda a trajetória do doutorado, uma irmã que me foi dada.
Ao time de Inteligência Artificial da Stilingue, Willian, JP, Valter, Dauberson, Avanço, Douglas, Evandro, Fernando, Amanda, Maxilene, Bruno e Dyovana, que acreditaram em mim e me deram a oportunidade de empregar tudo aqui aprendido.
A Deus que me guiou e deu força e fé para caminho
E ao meu filho Gael que está hoje em meu ventre, que me deu o respiro final nessa linda trajetória.

Resumo

O crescimento exponencial na geração e disponibilização de dados biológicos nas últimas décadas impulsionou o surgimento de bancos de dados como um recurso para orientar a inovação e a geração de novos *insights* biológicos. A ampla caracterização experimental desses dados é, em geral, inviável, dada a complexidade e escala desses, o que torna a classificação automática utilizando aprendizado de máquina uma alternativa essencial, mais rápida e barata. Muitos conjuntos de dados biológicos são de natureza hierárquica, com vários graus de complexidade, impondo diferentes desafios para se treinar, testar e validar modelos de classificação precisos e generalizáveis. Embora algumas abordagens para classificar dados hierárquicos tenham sido propostas, nenhuma orientação sobre sua utilidade, aplicabilidade e limitações foi explorada ou implementada até então. Isso inclui abordagens locais considerando a hierarquia, construindo modelos por nível ou nó, e global, usando uma abordagem de classificação plana. Para preencher essa lacuna, foi comparado sistematicamente o desempenho das abordagens *Local por Nível* e *Local por Nó* com uma abordagem *Global* aplicada a dois conjuntos de dados biológicos hierárquicos diferentes: *BioLiP* e *CATH*. Os resultados mostram como diferentes componentes de conjuntos de dados hierárquicos, como coeficiente de variação e previsão por profundidade, podem orientar a escolha de esquemas de classificação apropriados. Por fim, foram fornecidas diretrizes para apoiar esse processo ao embarcar em uma tarefa de classificação hierárquica, que ajudará a otimizar os recursos computacionais e o desempenho preditivo.

Palavras-chaves: Base de Dados Biológica. Hierarquia de Classes. Classificação Hierárquica. Predição de Função De Proteínas. Classificação Estrutural de Proteínas.

Abstract

The exponential growth in the generation and availability of biological data in recent decades has increased the importance of databases as a resource to guide innovation and the generation of new biological insights. The broad experimental characterization of these data is, in general, unfeasible, given their complexity and scale, which makes automatic data classification using Machine Learning an essential, faster, and cheaper alternative. Biological datasets are often hierarchical in nature, with varying degrees of complexity, imposing different challenges to train, test, and validate accurate and generalizable classification models. Although some approaches to classify hierarchical data have been proposed, no guidelines regarding their utility, applicability, and limitations have been explored or implemented. These include *Local* approaches considering the hierarchy, building models per level or node, and *Global* hierarchical classification, using a flat classification approach. To fill this gap, here we have systematically contrasted the performance of *Local per Level* and *Local per Node* approaches with a *Global* approach applied to two different hierarchical datasets: *BioLiP* and *CATH*. The results show how different components of hierarchical datasets, such as variation coefficient and prediction by depth can guide the choice of appropriate classification schemes. Finally, we provide guidelines to support this process when embarking on a hierarchical classification task, which will help optimize computational resources and predictive performance.

Keywords: Biological Database. Class Hierarchy, Hierarchical Classification. Protein Function Prediction. Protein Structural Classification.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

DAG	Grafo Acíclico Dirigido
MPL	Múltiplos Caminhos de Rótulos
SPL	Único Caminho de Rótulos
PD	Rotulagem de Profundidade Parcial
FD	Rotulagem de Profundidade Total
EC	Números de Classificação De Enzimas
V	Coeficiente de Variação
STD	Desvio Padrão
MCC	Coeficiente de Correlação de Matthew
CSM	*Cutoff Scanning Matrix*
AUC	Área Abaixo da Curva ROC
hP	Precisão Hierárquica
hR	*Hierarchical Recall*
hF	*F-Score* Hierárquico
DT	*Decision Tree*
RF	*Random Forest*
ET	*Extra Trees*
BL ACC	Acurácia Balanceada

SUMÁRIO

1. INTRODUÇÃO

Bancos de dados biológicos desempenham um papel importante na pesquisa contemporânea, fornecendo conjuntos de dados com curadoria e anotações para uso em vários campos, incluindo medicina, química e biotecnologia. Os dados biológicos geralmente são depositados por pesquisadores, coletados da literatura ou derivados de análises computacionais. Como essa recuperação de informações envolve intervenção humana, é passível de erros e, considerando a quantidade de dados biológicos coletados rotineiramente (ATTWOOD et al., 2011) a classificação automática de dados é muitas vezes necessária para alavancar a riqueza de informações desses repositórios.

Muitos conjuntos de dados biológicos têm uma natureza hierárquica, o que significa que eles têm classes (ou rótulos) que podem ser divididos em outras classes, como taxonomia do organismo (PRUESSE et al., 2007; SÖHNGEN et al., 2016), domínios estruturais de proteínas (MURZIN et al., 1995; PEARL et al., 2003; PIRES et al., 2011; SANDARUWAN; WANNIGE, 2021), vias metabólicas (KULMANOV et al., 2021; OGATA et al., 1999), classificações de enzimas (KULMANOV et al., 2021; STRODTHOFF et al., 2020), entre outros. Ao contrário das classificações planas onde classes são consideradas não relacionadas e independentes, as classificações hierárquicas associam rótulos a diferentes níveis de classificação (KOSMOPOULOS et al., 2015). Essas hierarquias se tornam um desafio para os algoritmos de classificação tradicionais, pois, em geral, não estão bem equipados para lidar com problemas de grande escala envolvendo centenas de milhares de classes hierarquicamente relacionadas, que são o caso de conjuntos de dados biológicos reais (KOSMOPOULOS et al., 2015). Por exemplo, uma representação taxonômica do domínio archaea possui cerca de 32.000 classes por nível, um grau de complexidade que levou a inconsistências de classificação comuns (BALVOČIŪTĖ; HUSON, 2017; PARKS et al., 2018).

Esforços anteriores mostraram que classificadores adaptados a esse tipo de dados hierárquicos complexos melhoram a recuperação de informações de forma eficaz (BREITWIESER; LU; SALZBERG, 2019; PARKS et al., 2018; ZHANG; WANG; WANG, 2017). Silla e Freitas, por exemplo, descrevem os principais desafios em tarefas de classificação hierárquica, incluindo desequilíbrio de classes e alto número de classes, previsão por profundidade e classificação em níveis mais profundos (ZHANG; WANG; WANG, 2017).

Ao longo dos anos, muitos métodos de classificação hierárquica foram propostos, incluindo novas métricas de avaliação (ZHANG; WANG; WANG, 2017) e abordagens de aprendizado profundo (CERRI; BARROS; DE CARVALHO, 2014; KOWSARI et al., 2017). No entanto, eles têm sido aplicados principalmente a problemas de classificação de texto (CERRI; BARROS; DE CARVALHO, 2014; KOSMOPOULOS et al., 2015; KOWSARI et al., 2017; SILLA; FREITAS, 2011), com pouco trabalho dedicado a enfrentar os desafios da classificação hierárquica em bancos de dados biológicos. Além disso, no campo da Bioinformática, *frameworks* hierárquicos têm sido usados para domínios específicos de aplicação (HENDERSON et al., 2019; NAKANO et al., 2017; PANTA et al., 2021; PYBUS et al., 2015; XIONG; ZENG; GONG, 2017); no entanto, eles compreendem base de dados taxonômicas, que apresentam limitações significativas, principalmente devido ao seu baixo nível de curadoria e qualidade dos dados (BALVOČIŪTĖ; HUSON, 2017; BEIKO, 2015; DESANTIS et al., 2006; PARKS et al., 2018; YILMAZ et al., 2014; YOON et al., 2017). Pouco, portanto, foi feito para avaliar de forma abrangente a utilidade, aplicabilidade e limitações de diferentes abordagens de classificação hierárquica aplicadas a diferentes bancos de dados biológicos (WEI et al., 2017). Neste trabalho, foram avaliadas as abordagens propostas por Silla e Freitas (ZHANG; WANG; WANG, 2017), e aplicadas a diferentes bancos de dados biológicos para investigar seus prós e contras e estabelecer diretrizes gerais de prática.

CATH é um banco de dados que mapeia as relações evolutivas em domínios de proteínas, que são classificados em quatro níveis: classe, arquitetura, topologia e superfamília homóloga (ORENGO et al., 1997). Os principais desafios de classificação relacionados ao *CATH* incluem um alto número de classes em níveis profundos, rotulagem de profundidade total e a natureza altamente desequilibrada das classes. BioLiP, por outro lado, é um banco de dados de dados de interação ligante-proteína (YANG; ROY; ZHANG, 2013). A partir deste banco de dados, extraímos a classificação de enzimas como um *proxy* para a função catalítica de proteínas, expressa como um número de Comissão Enzimática (EC). O BioLiP, ao contrário do CATH, não aceita rotulagem de profundidade total, mas apresenta classes altamente desequilibradas.

Neste estudo, foi avaliado e comparado, pela primeira vez, o desempenho de três abordagens de classificação hierárquica (Global, Local por Nó (Nó) e Local por Nível (Nível)) para dois conjuntos de dados. Por fim foram fornecidas diretrizes para escolher

estratégias apropriadas para classificar conjuntos de dados hierárquicos considerando suas principais características.

Os próximos tópicos estão divididos em três seções: Classificação Hierárquica (seção 1.1), Desafios na Classificação Hierárquica (seção 1.2) e Abordagens de Classificação Hierárquica (seção 1.3). Os tipos e características das topologias hierárquicas são tratados na seção 1.1, a seção Desafio na Classificação Hierárquica aborda os quatro tipos de desafios desse modelo de classificação: Previsão por profundidade, Níveis Profundos de Classificação, Classes não Balanceadas e Número de Classes; e a seção 1.3 aprofunda-se nos tipos de abordagens de classificação hierárquica: Abordagem Global, Abordagem Local por Nível e Abordagem Local por Nó.

1.1. CLASSIFICAÇÃO HIERÁRQUICA

Na classificação tradicional ou plana, um modelo é treinado para atribuir cada objeto a uma única classe pertencente a um número finito de classes. Porém, quando o objeto está associado a diferentes níveis de classificação, há uma especialização dessa tarefa, denominada classificação hierárquica.

Uma classificação hierárquica pode ser organizada como uma topologia de árvore ou de grafo acíclico dirigido (DAG). Em uma topologia de árvore, cada classe-filho está associada a uma única classe-pai ou ancestral (Figura 1A), enquanto em uma topologia de DAG cada classe-filho pode ser associada a uma ou mais classes-pai (Figura 1B) (KOSMOPOULOS et al., 2015). A principal diferença entre as topologias está no resultado da classificação: enquanto em uma árvore existe um único caminho para classificar cada nó folha, em um DAG pode haver mais de um.

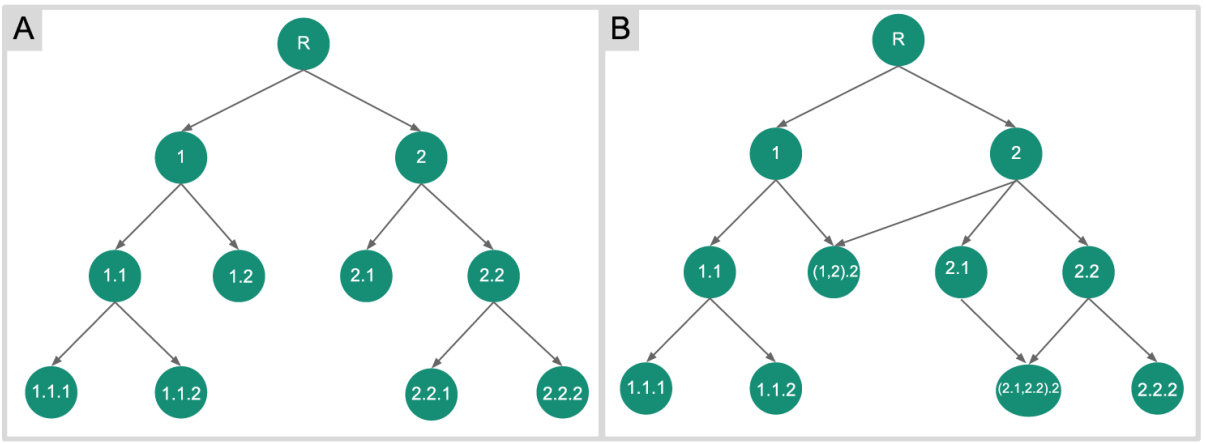

Figura 1. Exemplos de topologias hierárquicas entre Árvore e DAG (A) Árvore: um grafo não direcionado em que quaisquer arestas são conectadas por exatamente um nó e (B) Grafo Acíclico

Direcionado (DAG): Um grafo direto sem ciclos direcionados. As arestas do DAG só seguem uma direção, no entanto, um nó pode ser conectado por várias arestas.

Classificações hierárquicas podem ser categorizadas com base em três características principais (SILLA; FREITAS, 2011) :

- Tipo de hierarquia, em que as classes são organizadas (Árvore ou DAG);
- Classificação de rótulo único ou multi-rótulo (*ou seja*, permitindo que os pontos de dados sigam vários caminhos de classificação);
- Com base na profundidade de rotulagem de dados; ou seja, todas as instâncias têm rótulos até os nós folha, que representam os níveis mais profundos em uma hierarquia, ou rotulagem de profundidade parcial.

1.2. DESAFIOS NA CLASSIFICAÇÃO HIERÁRQUICA

Considerando X como os espaços de instâncias, um problema de classificação hierárquica consiste em encontrar uma função (classificador) f para mapear cada instância $x_i \in X$ para um conjunto de classes $C_i \in C$, com C sendo o conjunto de classes do problema. A função f deve respeitar as restrições da hierarquia e otimizar um critério de qualidade (CERRI et al., 2016). Como restrições da hierarquia, quando uma classe é prevista, todas as suas superclasses também devem ser previstas automaticamente.

Problemas de classificação hierárquica são formalmente definidos como tuplas (γ, ψ, ϕ), onde γ especifica a topologia (Árvore ou DAG), ψ descreve se as instâncias são classificadas em múltiplos caminhos de rótulos (MPL) ou em um único caminho de rótulos (SPL), e ϕ determina se a classificação pode parar em um nó interno da hierarquia (nó folha não obrigatório ou rotulagem de profundidade parcial (PD)), ou se deve continuar até que um nó folha seja alcançado (nó folha obrigatório ou rotulagem de profundidade total (FD)) (CERRI et al., 2016; SILLA; FREITAS, 2011).

À medida que o número de níveis em uma hierarquia cresce, a complexidade e o esforço necessários para alcançar uma previsão satisfatória aumentam. A previsão por profundidade pode ser usada com duas estratégias: rotulagem de profundidade total ou parcial (Figura 2A). A rotulagem completa é usada quando cada nó é classificado em todos os níveis da hierarquia, desde os nós raízes até os nós folhas. A desvantagem da rotulagem completa é que os dados recebem uma classificação independentemente da confiança da previsão. Por

outro lado, na rotulagem de profundidade parcial, a tarefa de previsão é interrompida quando a confiança é baixa, garantindo a confiabilidade da classificação. Além dos problemas relacionados à topologia, frequentemente observados em conjuntos de dados biológicos, dois outros desafios comuns podem gerar vieses nos modelos de classificação: classes desbalanceadas (algo bem conhecido para modelos de aprendizado de máquina, que tendem a privilegiar classes maiores) e elevado número de classes (que reduzem os limites das classes, dificultando o processo de aprendizagem).

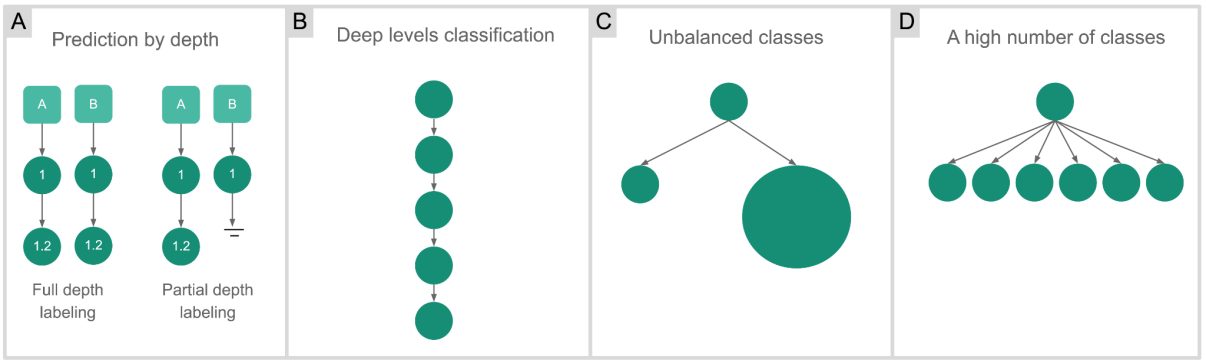

Figura 2. Representação dos desafios enfrentados na classificação hierárquica de dados. (A) Previsão por profundidade: descreve a profundidade do rótulo das instâncias de dados. A rotulagem completa indica que cada instância é rotulada com classes em todos os níveis, desde o primeiro nível até o nível folha. A rotulagem de profundidade parcial indica que pelo menos uma instância tem uma profundidade parcial de rotulagem, ou seja, o valor do rótulo da classe em algum nível é desconhecido. (B) Classificação de níveis profundos: A complexidade da classificação é afetada pelo número de níveis na topologia. (C) Classes desequilibradas: Há uma distribuição desigual de classes no conjunto de dados, o que pode penalizar as classes mais baixas no processo de classificação. (D) Grande número de classes: Um número expressivo de classes, principalmente no último nível, afeta a complexidade do modelo.

Em resumo, os quatro principais desafios na classificação hierárquica são (Figura 2):

A. **Previsão por profundidade:** As instâncias são classificadas até o último nível hierárquico ou até que a precisão da previsão seja suficiente.

B. **Níveis profundos de classificação:** quanto mais profundo for um nível de classificação, mais difícil será obter uma boa precisão de previsão.

C. **Classes não balanceadas:** Uma grande diferença no número de instâncias entre diferentes classes.

D. **O número de classes:** Um modelo preditivo para um grande número de classes precisa ser treinado.

1.3. ABORDAGENS DE CLASSIFICAÇÃO HIERÁRQUICA

Duas abordagens principais têm sido usadas para lidar com problemas de classificação hierárquica: locais e globais.

Classificações globais consideram os caminhos de classificação como um único rótulo, ou seja, a hierarquia de dados é desconsiderada e o classificador funciona como um classificador simples; em outras palavras, um único modelo preditivo é gerado para todos os níveis da hierarquia. Em contraste, as abordagens locais, que são divididas em baseadas por Nó e Nível, consideram a hierarquia de rótulos. A Figura 3 mostra as diferenças entre as abordagens, onde as linhas tracejadas são os modelos preditivos gerados.

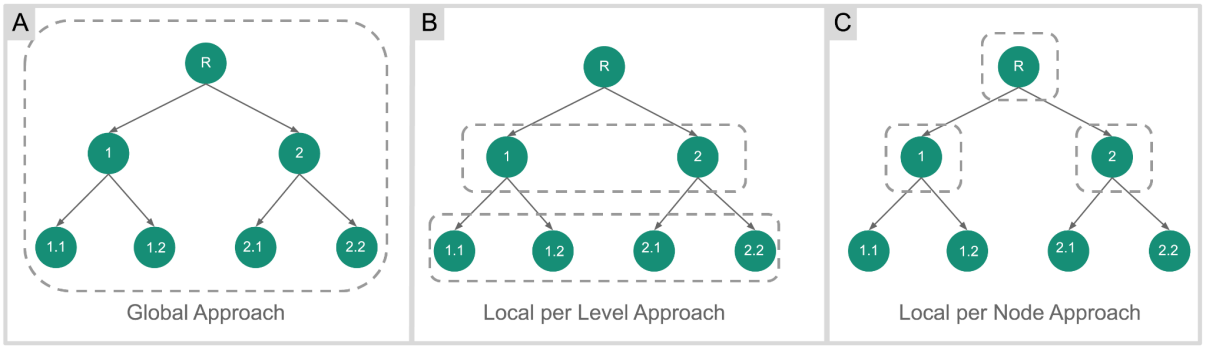

Figura 3. Abordagens de classificação hierárquica. (A) Abordagem Global: considera toda a hierarquia de classes de uma só vez. (B) Abordagem Local por Nível: consiste em treinar um classificador multiclasse para cada nível da hierarquia de classes. (C) Abordagem Local por Nó: consiste em treinar um classificador multiclasse para cada nó pai na hierarquia de classes. (Adaptado de Silla e Freitas (SILLA; FREITAS, 2011)).

Na abordagem por nível, um classificador é desenvolvido para cada nível da hierarquia considerando todos os nós de cada nível como uma classe. Considerando o exemplo da Figura 3B, dois classificadores seriam treinados, um para cada nível de classe para prever uma ou mais classes em seu nível de classe correspondente (SILLA; FREITAS, 2011).

A abordagem de classificação por nó consiste em desenvolver um classificador multiclasse para cada nó pai da hierarquia de classes (SILLA; FREITAS, 2011). Normalmente, a abordagem de classificação de nós segue uma predição obrigatória de nós folha, uma vez que esta abordagem associa um classificador multiclasse a cada nó interno da hierarquia. Portanto, cada nó aprende a diferenciar entre suas subclasses (NAKANO et al., 2017).

Em termos de número de modelos, o conjunto de rótulos para cada classificador na abordagem por nó será seus nós filhos. Nesta abordagem, temos menos classes por modelo em comparação com a abordagem global; no entanto, produz consideravelmente mais modelos (com menos informações por modelo disponível para treinamento). A abordagem por nível produz menos modelos do que a abordagem por nó, porém, gera modelos mais complexos dado o aumento do número de classes por manipular todo o nível hierárquico.

2. OBJETIVOS

2.1. OBJETIVOS GERAIS

O presente trabalho teve como objetivo avaliar e comparar o desempenho de três abordagens de classificação hierárquica para dois conjuntos de dados biológicos (Global, Local por Nó (Nó) e Local por Nível (Nível)). Gerando diretrizes para escolher estratégias apropriadas para classificar conjuntos de dados hierárquicos considerando suas principais características.

2.2. OBJETIVOS ESPECÍFICOS

1. Avaliar e escolher o melhor conjunto de características para cada base de dados;
2. Avaliar e escolher o melhor algoritmo classificador para o estudo;
3. Desenvolver uma nova metodologia para avaliação de diferentes estratégias de classificação hierárquica e sua relação com as propriedades inerentes às bases de dados empregadas;
4. Desenvolver abordagens de classificação capazes de lidar com desbalanceamento de classes, elevado número de classes, grande profundidade e classificação parcial.
5. Avaliar o desempenho das abordagens de classificação hierárquica aplicadas à classificação estrutural e funcional de proteínas.
6. Replicar a avaliação das abordagens de classificação hierárquica a outros contextos de classificações hierárquicas biológicas.
7. Definir diretrizes iniciais para escolha das classificações hierárquicas biológicas guiadas pelas características das bases: desbalanceamento de classes, elevado número de classes, grande profundidade e classificação parcial.

3. MATERIAIS E MÉTODOS

A metodologia proposta foi dividida em três etapas, como demonstrado na Figura 4. A primeira etapa do trabalho envolveu a seleção das bases de dados biológicas que tem como critério de seleção os desafios citados anteriormente e a aplicabilidade em diferentes áreas biológicas. Para cada base selecionada foram avaliadas diversas abordagens de classificação hierárquica e seus resultados contrastados com o seu perfil de base. A partir da escolha da melhor abordagem, em trabalhos futuros, serão treinados novos classificadores baseando-se em técnicas de aprendizado de máquina, os quais serão disponibilizados para novos estudos através de servidores web. Por fim, esses modelos serão utilizados para identificar erros nas bases selecionadas a fim de gerar bases mais curadas e com maior confiabilidade.

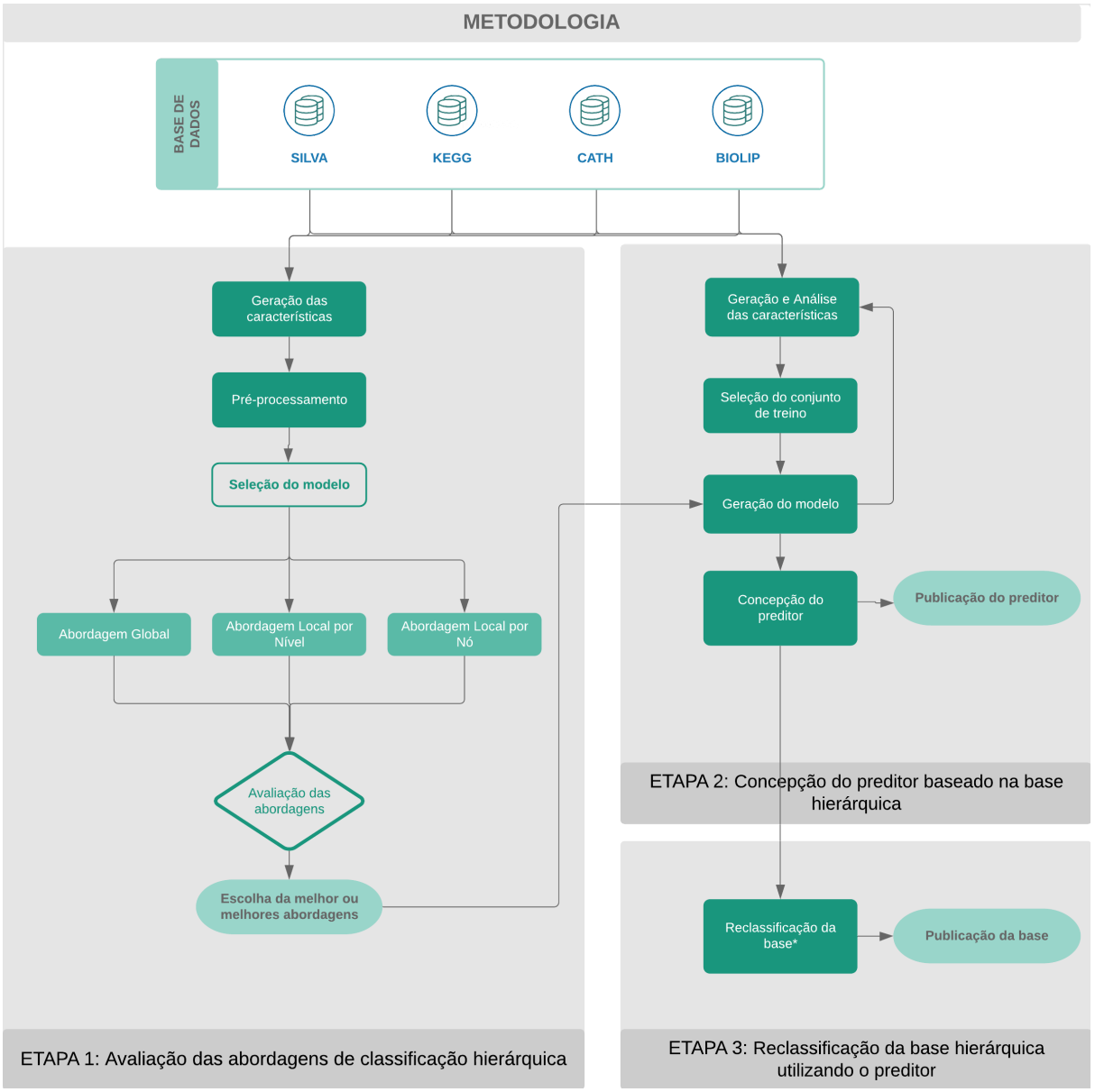

Figura 4. Visão geral da metodologia proposta no trabalho. A metodologia foi dividida em três etapas. A primeira etapa de avaliação das diferentes abordagens de classificação hierárquica, a segunda etapa com geração dos preditores baseados na melhor abordagem, e por fim a publicação das bases reclassificadas (trabalhos futuros).

Porém, conforme o desenvolvimento do trabalho foi avançando, os desafios enfrentados demonstraram-se maiores que o previsto. Portanto, para esse trabalho, foi desenvolvido e apresentado os resultados da primeira etapa, Avaliação das abordagens de classificação hierárquica. Esta etapa se dividiu em 4 fases: seleção de conjuntos de dados, seleção de algoritmos, análises exploratórias, engenharia de características e por fim a análise

de desempenho de abordagens hierárquicas. A primeira fase do trabalho (seção 3.1) envolve a seleção dos conjuntos de dados para a base de dados *BioLiP* e *CATH*. Também foram exploradas outras bases de dados biológicos a fim de conectar os resultados obtidos das bases de dados selecionadas. Foi aplicada também a engenharia de características em cada uma das bases, baseada em metodologias da literatura (seção 3.4). Para as análises hierárquicas foram feitas a seleção dos algoritmos de classificação (seção 3.2), a partir de 3 que apresentaram o melhor resultado. Cada base selecionada foi submetida à análise de desempenho das abordagens de classificação hierárquica (seção 3.5) e seus resultados contrastados com o seu perfil de base feito na etapa de análise exploratória (seção 3.3).

3.1. SELEÇÃO DO CONJUNTO DE DADOS

A avaliação descrita neste estudo se concentrou em quatro bancos de dados hierárquicos *BioLiP, CATH, SILVA e KEGG*. As bases selecionadas são descritas a seguir:

(a) ***BioLiP, Enzyme Commission numbers***: compreende em uma base curada de interações de ligação ligante-proteína (YANG; ROY; ZHANG, 2013). A partir dessa base, a classificação de enzimas em sua função catalítica, expressa na forma de número EC (*Enzyme Commission numbers*) foi extraída e utilizada. Esta possui 4 níveis e é baseada nas reações químicas que as enzimas catalisam. Esta base de dados contém classes desbalanceadas e aceita a previsão incompleta dos dados.

(b) ***CATH:*** consiste em uma base de dados que apresenta relações evolutivas dos domínios proteicos. Tais domínios são classificados em 4 níveis: classe, arquitetura, topologia e superfamília homóloga (DAS et al., 2021). *CATH* é uma base curada, entretanto, apresenta desafios como número elevado de classes em níveis mais profundos, classes desbalanceadas e aceita a previsão parcial na classificação.

(c) ***SILVA, high quality ribosomal RNA Database:*** consiste em uma base de dados de classificação taxonômica de organismos que fornece sequências de pequenos RNAs ribossômicos (16S / 18S, SSU) e grandes subunidades (23S / 28S, LSU) para todos os três domínios da vida (Bacteria, Archaea e Eukarya) (PRUESSE et al., 2007). Foi avaliada neste projeto a classificação de pequenos RNAs ribossomais dos domínios Archea e Bacteria. Essa classificação é feita em 7 níveis: domínio, filo, classe, ordem, família, gênero e espécies. Esta

base além de não ser uma base de dados curada, apresenta os desafios de número elevado de classes, o desbalanceamento entre elas e aceita apenas a previsão completa dos dados.

(d) ***KEGG, Kyoto Encyclopedia of Genes and Genomes:*** é uma base de dados de interação entre funções de alto nível entre sistemas biológicos (KANEHISA et al., 2017). Neste projeto foi utilizado o KEGG BRITE, classificação funcional de entidades biológicas, o qual incorpora diferentes tipos de relações: genes e proteínas, compostos e reações, drogas, doenças, células e organismos. Ela contém classes desbalanceadas, além de um número elevado de quantidade de classes, aceita a previsão incompleta, e não é uma base de dados altamente curada.

Tais bases foram selecionadas por múltiplos motivos. Em primeiro lugar, sua estrutura representa os desafios enfrentados pela classificação hierárquica (representada na Figura 2). *CATH* apresenta dois desafios para classificação: um grande número de classes, e classes desbalanceadas, com um esquema de rotulagem de profundidade total, enquanto o *BioLiP* apresenta rotulagem de profundidade parcial e classes desbalanceadas como seus principais desafios. Além disso, essas base de dados foram escolhidos por sua popularidade e seu alto nível de curadoria e qualidade dos dados, que limitam os fatores de confusão na análise e minimizam os erros de classificação. Avaliamos bancos de dados hierárquicos biológicos alternativos disponíveis publicamente (Tabela 1 e Figura 5) e identificamos alternativas potenciais (*por exemplo*, bancos de dados do *Silva* (PRUESSE et al., 2007) e *KEGG* (OGATA et al., 1999)). No entanto, não apresentavam o mesmo nível de curadoria das demais bases, nem pertenciam a domínios de aplicações já contemplados pelo *CATH* e *BioLiP,* portanto não foram utilizados extensivamente.

Tabela 1. Revisão dos bancos de dados hierárquicos biológicos disponíveis publicamente mais usados

Base de dados	**Domínio**	**#Citações ***
BioLiP (YANG; ROY; ZHANG, 2013)	Protein function classification	463
CATH (DAS et al., 2021; DAWSON et al., 2017; ORENGO et al., 1997; PEARL et al., 2003)	Protein structure classification	3128
Pfam (BATEMAN et al., 2000)	Protein domain classification	2877
GreenGenes (DESANTIS et al., 2006)	Taxonomic classification	9759
KEGG Brite - EC Number (KANEHISA et al., 2017)	Protein function classification	4791
NCBI taxonomy (FEDERHEN, 2012)	Taxonomic classification	1072
OTT (HINCHLIFF et al., 2015)	Taxonomic classification	523
RDP (COLE et al., 2014)	Taxonomic classification	3265
Silva (QUAST et al., 2013)	Taxonomic classification	1651

* Citação do Google Acadêmico em Março 2022.

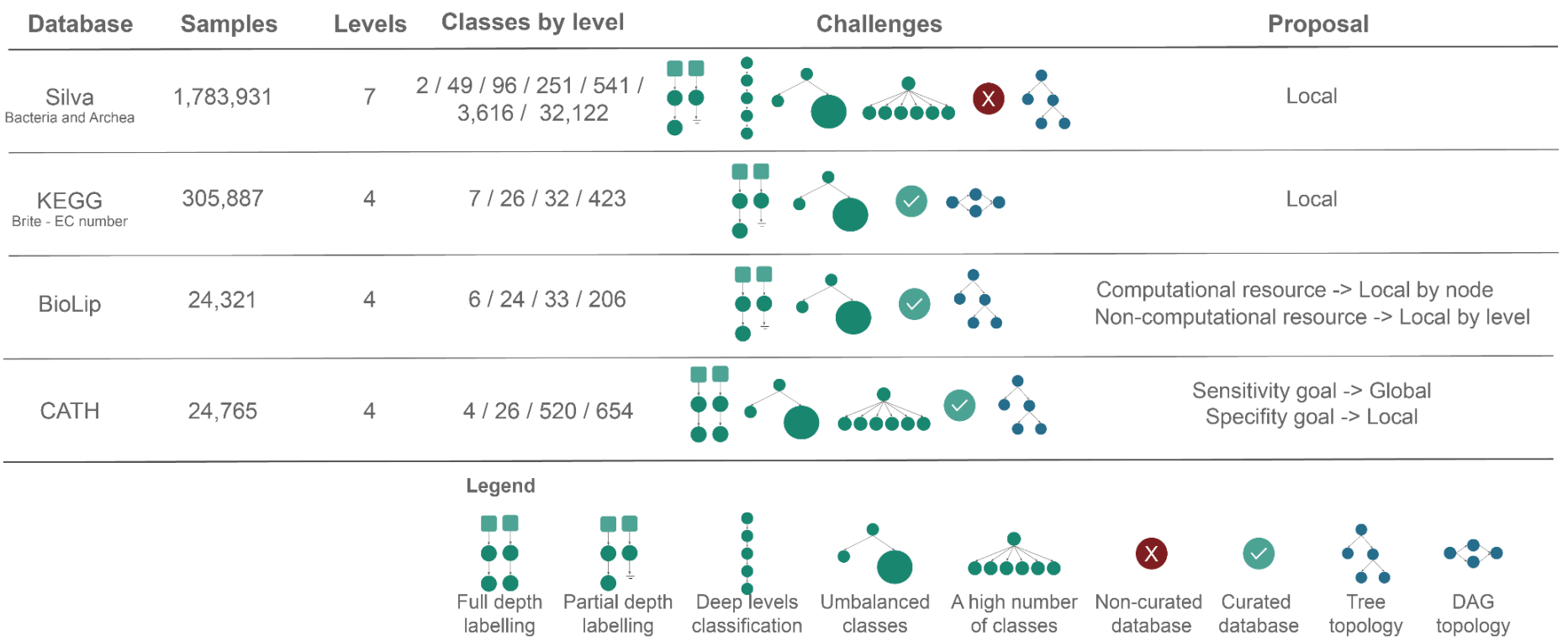

Figura 5. Análises exploratórias para diferentes conjuntos de dados biológicos hierárquicos. As análises foram realizadas utilizando Silva (Bactéria e Archaea), *KEGG* (Brite - número CE), *BioLiP* e *CATH* para entender as principais características de conjuntos de dados biológicos hierárquicos de diferentes domínios. Avaliamos números estatísticos para cada conjunto de dados, quantidade de amostras (amostras), quantidade de níveis na hierarquia (níveis) e quantidade de classes presentes em cada nível (classes por nível), desafios e a proposta. Os desafios referem-se aos problemas enfrentados pela classificação hierárquica, curadoria de banco de dados e topologia. Os problemas enfrentados pela classificação hierárquica geralmente são: Previsão por profundidade (rotulagem de profundidade total e rotulagem de profundidade parcial), classificação de níveis profundos, classes não balanceadas e um número alto de classes. Para a curadoria de banco de dados, usamos uma classificação binária: não curado ou curado, e para a topologia, os conjuntos de dados seguem uma árvore ou DAG. Na coluna Proposta, sugerimos abordagens que podem ser usadas para cada banco de dados aplicando nosso Note que para Silva e *KEGG* estamos generalizando os resultados dos experimentos realizados neste trabalho, com base em uma análise exploratória realizada com ambos os conjuntos de dados.

A preparação do conjunto de dados do *CATH* foi através do download da base via website, incluindo todos os domínios de proteínas com dados estruturais disponíveis no Protein Data Bank (PDB) com suas respectivas classificações. As características foram geradas com base no algoritmo CSM (PIRES et al., 2011). Para a base do *BioLiP* os números de classificação de enzimas (números EC) foram recuperados do website e os arquivos FASTA com as sequências de proteínas, baixados e apresentados ao *iFeatures* (CHEN et al., 2018) para geração de recursos.

3.2. SELEÇÃO DE ALGORITMOS CLASSIFICADORES

Para priorizar os algoritmos de aprendizado usados nos experimentos, realizamos a seleção de modelos via abordagem global em *CATH*, *BioLiP* e *Silva*, usando validação cruzada de 10-*folds,* quantidade de *folds* mais popualr utilizado nos estudos, para seis algoritmos diferentes (Tabela 2):

(a) **XGBoost:** Baseado em Árvore de Decisão com Aumento de Gradiente (Anon n.d.). O princípio do Aumento de Gradiente é a capacidade de combinar resultados de muitos classificadores fraco, como árvores de decisão, que se combinam para formar algo parecido com um comitê forte de decisão.

(b) ***Random Forest:*** É criada a partir da combinação de Árvores de Decisão, ou seja, uma floresta, na qual cada árvore é treinada sobre várias subamostras aleatória dos dados (BREIMAN, 2001). Diferente das árvores de decisão simples, o Random Forest busca a melhor característica em um subconjunto aleatório de características. A combinação de modelos cria uma diversidade e diminui a variância em comparação a uma árvore de decisão única, melhorando assim o resultado geral e diminuindo o sobreajuste de dados.

(c) **Decision Trees:** É um método de aprendizagem supervisionada, não paramétrico, que usa Árvores Binárias em problemas de classificação e regressão. Seu objetivo é criar um modelo para predizer o valor de uma variável alvo através de regras de decisão simples, inferidas a partir do conjunto de características (BREIMAN et al., 1984) Uma árvore de decisão, por ter apenas um caminho a seguir, é conhecida por ter uma alta variância, pois geralmente sobre ajusta os dados.

(d) **Extremely Randomized Trees:** Assim como o Random Forest, o Extremely Randomized Trees, também conhecido como Extra Trees, é um método de combinação de métodos de aprendizagem, baseado em árvores de decisão (GEURTS;

ERNST; WEHENKEL, 2006). Entretanto, ele apresenta duas principais diferenças: não faz reamostragem e as divisões são aleatórias. Logo, no Extra Trees a aleatoriedade não vem da reamostragem dos dados como o Random Forest, mas sim da das divisões aleatórias dos nós das árvores, causando uma baixa variância.

(e) **Radius Neighbors:** Este classificador é uma extensão do classificador Nearest Neighbor Search (Chazelle 1983), baseado em regressão múltipla, onde o vizinho mais próximo é determinado pela distância hiperparâmetro de raio. Tal classificador é mais utilizado para dados esparsos, mais adequado para se utilizar em conjunto de dados que contem um número maior de características, e também quando a multicolinearidade é experimentada em conjunto. Multicolineridade é a existência da correlação entre variáveis independentes e dados modelados.

(f) **Ridge:** Este classificador também é baseado em métodos de regressão, assim como Radius Neighbors (Hazarika et al. 2021). Ele penaliza os coeficientes acrescentando na função de custo do modelo para evitar overfitting. Isso é benéfico quando se tem poucos dados de treinamento disponíveis, no entanto, se o prazo de penalidade for muito grande, pode resultar em underfitting.

Tabela 2. Seleção do modelo com as bases de dados hierárquicas (*CATH*, *BioLiP* e *Silva*) usando validação cruzada com *10 dobras*.

Modelos	**Melhor acurácia**
XGBoost	0.60
Random Forest	0.86
Decision Tree	0.79
Extra Trees	0.78
Radius Neighbors	0.08
Ridge	0.49

Decision Tree, Random Forest e *Extra Trees* tiveram os melhores desempenhos. Assim para a seleção da melhor abordagem eles foram comparados.

3.3. ANÁLISES EXPLORATÓRIAS

Nossa análise utilizou abordagens de classificação hierárquica definidas na literatura para avaliar sua utilidade e aplicabilidade (SILLA; FREITAS, 2011). Os algoritmos de classificação utilizados foram implementados em Python, utilizando a biblioteca Scikit-learn (PEDREGOSA et al., 2011). A configuração das máquinas utilizadas para realizar os experimentos está descrita na Figura 6.

Análises locais foram realizadas em Intel(R) Xeon(R) CPU E7-4850 @ 2.00GHz, 80 cores, 65.78 GB de memória, CentOS Linux 7 (Core).
Análises globais foram realizadas Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30GHz, 48 cores, 395.984 GB de memória, CentOS Linux 7 (Core).

Figura 6. Configuração das máquinas utilizadas para os experimentos.

Versões gratuitas do *CATH*, *BioLiP e Silva* foram baixadas (Figura 7). O problema de classificação hierárquica para *CATH* é descrito como $\gamma = tree$, $\psi = SPL$, $\phi = FD$, no *BioLiP*, é descrito como $\gamma = tree$, $\psi = SPL$, $\phi = PD$, e no *Silva* é descrito como $\gamma = tree$, $\psi = SPL$, $\phi = PD$. As bases de dados *CATH e BioLiP* contém quatro níveis de classificação, enquanto a base *Silva* possui 7 níveis, e em cada nível podemos observar uma representação de classe desbalanceada.

As análises exploratórias iniciais foram feitas na base *Silva.* Para tais análises foram considerados apenas dois domínios, *Archaea* e *Bacteria*. Olhando a diferença entre os níveis (Anexo A - Figura A1) já é possível notar um desbalanceamento comparando os 7 níveis. Nos primeiros níveis de classificação, domínio, filo e classe, foi possível identificar um desbalanceamento entre os rótulos de classificação (Anexo A - Figura A2, A3 e A4). Nos níveis mais profundos a maioria das classes demonstrou possuir até 50 sequências no nível de Gênero (Anexo A - Figura A7) e 5 sequências a nível de Espécie (Anexo A - Figura A8). Observa-se que quanto mais profundo, mais complexo se torna o trabalho de classificação devido a quantidade de sequências por classes.

Outro fator que também foi encontrado na base de dados do Silva foi a não curadoria dos dados. A nível de Família, 10.636 sequências classificadas são desconhecidas ou incertas (Anexo A - Figura A9), enquanto a nível de gênero esse valor aumenta para 90.759 sequências (Anexo A - Figura A10), e a nível de espécie para 1.345.239, o que corresponde a 78% de toda a base (Anexo A - Figura A10). Essas observações demonstram que além da não curadoria dos dados essa base também enfrenta a rotulagem de profundidade parcial. Levando em conta os problemas de curadoria, profundidade de classificação, desbalanceamento das

classes e profundidade parcial, optamos por incluir nas análises de avaliação as bases do *CATH* e *BioLiP,* devido a menor complexidade, e apenas elas serão citadas a partir de agora.

De uma perspectiva de cima para baixo, na Figura 8, é possível observar a distribuição das classes em ambos os conjuntos de dados. A divisão dos agrupamentos segundo a classificação dos dados, mostra o primeiro e segundo níveis da hierarquia da taxonomia das bases de dados, destacando a natureza altamente desbalanceada e o número variável de classes nos diferentes níveis.

CATH: version 4.2
BioLiP: version 2019-07-15
Silva: version 128
KEGG: version 2022-03-10

Figura 7. Versões usadas das bases de dados.

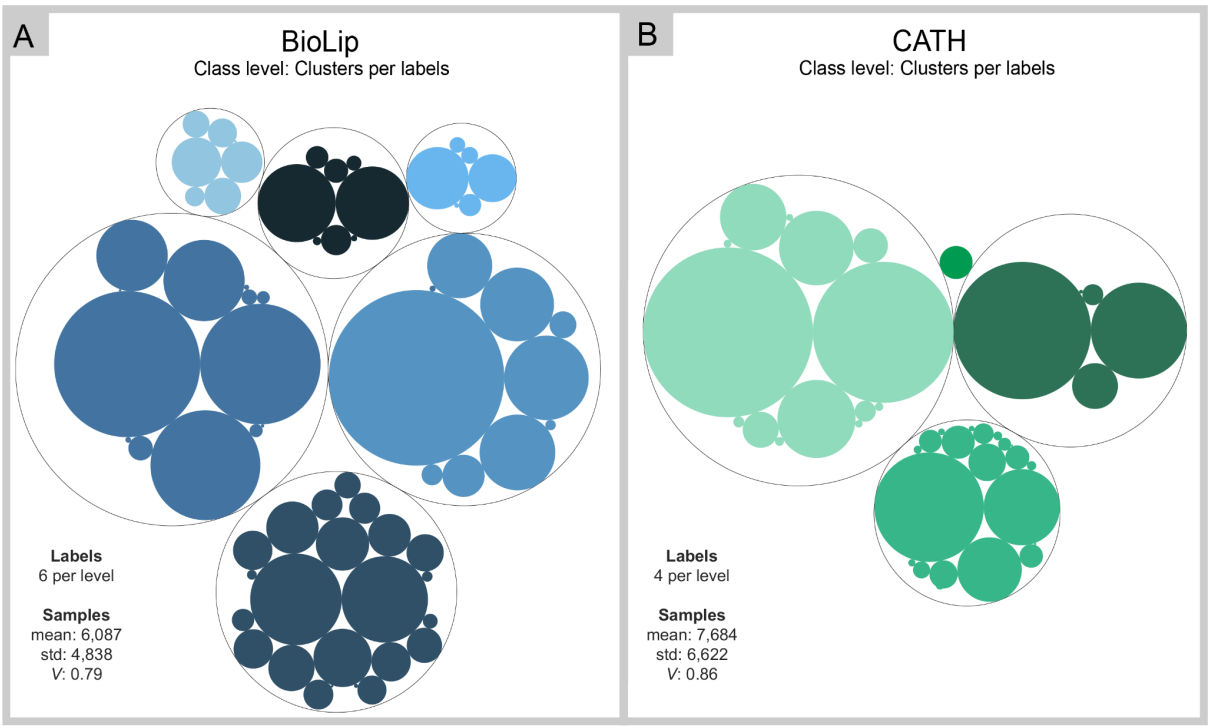

Figura 8. Distribuição de entradas em cada classe de *BioLiP* (A) e *CATH* (B) de uma perspectiva de cima para baixo. A divisão de agrupamentos demonstra o primeiro nível de cada uma das classes e o tamanho dos círculos apresenta o número de amostras para cada classe.

Em relação à topologia, à medida que descemos no nível da árvore, ambas apresentam um crescimento exponencial do número de classes ou rótulos (Tabela 3 e Anexo B). No último nível, há uma média de nove classes por nó para o *CATH* e 23 para *BioLiP*, com uma

representação média de 46 e 177 amostras por classe, respectivamente. A disparidade entre a média e o desvio padrão (*STD*) dos dados indica uma alta dispersão no conjunto de dados. A razão entre essas medidas, denominada como Coeficiente de Variação (*V*), mostra o grau de variação das amostras em cada nível. *V* é amplamente utilizado para medir a dispersão de dados ou para avaliar problemas em resultados de experimentos (BEDEIAN; MOSSHOLDER, 2000; BROWN, 1998; REED; LYNN; MEADE, 2002). *V* indica quão grandes as diferenças dentro do grupo tendem a ser em comparação com sua média. O limiar utilizado para avaliar a dispersão de um conjunto varia de acordo com o domínio. No entanto, em termos de distribuição estatística, o *STD* de uma distribuição exponencial é igual à sua média, então seu *V* é igual a 1. Distribuições com $V < 1$ são consideradas de baixa variância, enquanto aquelas com $V > 1$ são consideradas de alta variação (TIAN, 2005).

Tabela 3. Caracterização das bases de dados em relação a número de classes e amostras por nível.

		1º Level		**2º Level**		**3º Level**		**4º Level**	
		CATH	*BioLiP*	*CATH*	*BioLiP*	*CATH*	*BioLiP*	*CATH*	*BioLiP*
Classes	por nível	4	6	26	23	520	32	654	206
	por nó	-	-	10	11	46	6	9	23
Amostras	média	7,684.25	6,087.67	1,182.19	1,578.52	59.00	1,121.69	46.00	177.31
	STD	6,622.29	4,838.16	2,310.56	2,351.38	291.00	2,868.25	509.00	532.99
	V	0.86	0.79	1.95	1.49	4.93	2.56	11.07	3.01

A Tabela 3 mostra que, no primeiro nível, *CATH* e *BioLiP* têm quase a mesma variação, que é inferior a 1. No segundo nível, esse valor aumenta de forma consistente, principalmente em *CATH*. No terceiro nível as diferenças entre os conjuntos de dados tornam-se mais proeminentes, com o STD para o último nível de *CATH* sendo 11 vezes maior que sua média, ao contrário do STD do *BioLiP*, que é três vezes maior. Essas características reforçam os desafios de classificação que apresentamos anteriormente: um grande número de classes para um mesmo nó, uma representação desbalanceada dessas classes e a dificuldade de classificar os nós à medida que o nível se aprofunda.

Em relação ao problema de previsão por profundidade, no *BioLiP* uma amostra nem sempre é anotada até o último nível, permitindo uma marcação parcial de profundidade. Portanto, se a classe dos últimos níveis estiver indisponível ou duplicada, a última classificação é considerada uma folha.

3.4. ENGENHARIA DE CARACTERÍSTICAS

Usamos descritores de composição de aminoácidos de *iFeature (CHEN et al., 2018)* para rA Tabela 3 mostra que, no primeiro nível, CATH e BioLiP têm quase a mesma variação, que é inferior a 1. No segundo nível, esse valor aumenta de forma consistente, principalmente em CATH. No terceiro nível as diferenças entre os conjuntos de dados tornam-se mais proeminentes, com o STD para o último nível de CATH sendo 11 vezes maior que sua média, ao contrário do STD do BioLiP, que é três vezes maior. Essas características reforçam os desafios de classificação que apresentamos anteriormente: um grande número de classes para um mesmo nó, uma representação desbalanceada dessas classes e a dificuldade de classificar os nós à medida que o nível se aprofunda.

Em relação ao problema de previsão por profundidade, no BioLiP uma amostra nem sempre é anotada até o último nível, permitindo uma marcação parcial de profundidade. Portanto, se a classe dos últimos níveis estiver indisponível ou duplicada, a última classificação é considerada uma folha.epresentar proteínas em *BioLiP*, pois estes foram amplamente utilizados em trabalhos anteriores de modelagem de informações neste banco de dados (AKCESME, 2015; SHEN; TANG; GUO, 2019; SONG et al., 2019). Assinaturas baseadas em grafos foram usadas para representar estruturas de proteínas em *CATH*, usadas também anteriormente para modelar estrutura e função de proteínas (DA SILVA et al., 2022; DA SILVEIRA et al., 2009; PIRES; ASCHER, 2016; PIRES et al., 2013), prever efeitos de mutação (MYUNG et al., 2020; PIRES; ASCHER; BLUNDELL, 2014a, 2014b; PIRES; ASCHER, 2017; PIRES; BLUNDELL; ASCHER, 2016; PIRES; RODRIGUES; ASCHER, 2020; RODRIGUES; ASCHER; PIRES, 2018) e de modelar a hierarquia do *CATH* (PIRES et al., 2011). Neste estudo, escolhemos descritores conhecidos e validados para cada banco de dados, uma vez que avaliar conjuntos de descritores ideais estava fora do escopo do presente trabalho.

Valores de *Shapley* (LUNDBERG; LEE, 2017) são usados para simplificar os modelos preditivos e reduzir os requisitos de tempo computacional, eles explicam a contribuição das

características em todas as combinações possíveis de um modelo supervisionado. Para o *BioLiP*, foram selecionados as 60 melhores características classificadas pelo valor de *Shapley* e avaliamos o modelo usando o Coeficiente de Correlação de Matthew (*MCC*) e Revocação. Não observamos alteração nas métricas após a variação do número de características (Figura 9), o que mostra que usar 60 foi o suficiente para avaliar as abordagens (Figura 10). Para o *CATH*, 10 características foram selecionadas pela distância entre os carbonos α (a lista final das feições utilizadas pode ser encontrada no Anexo C). Uma descrição detalhada dos procedimentos de seleção das característica está disponível abaixo.

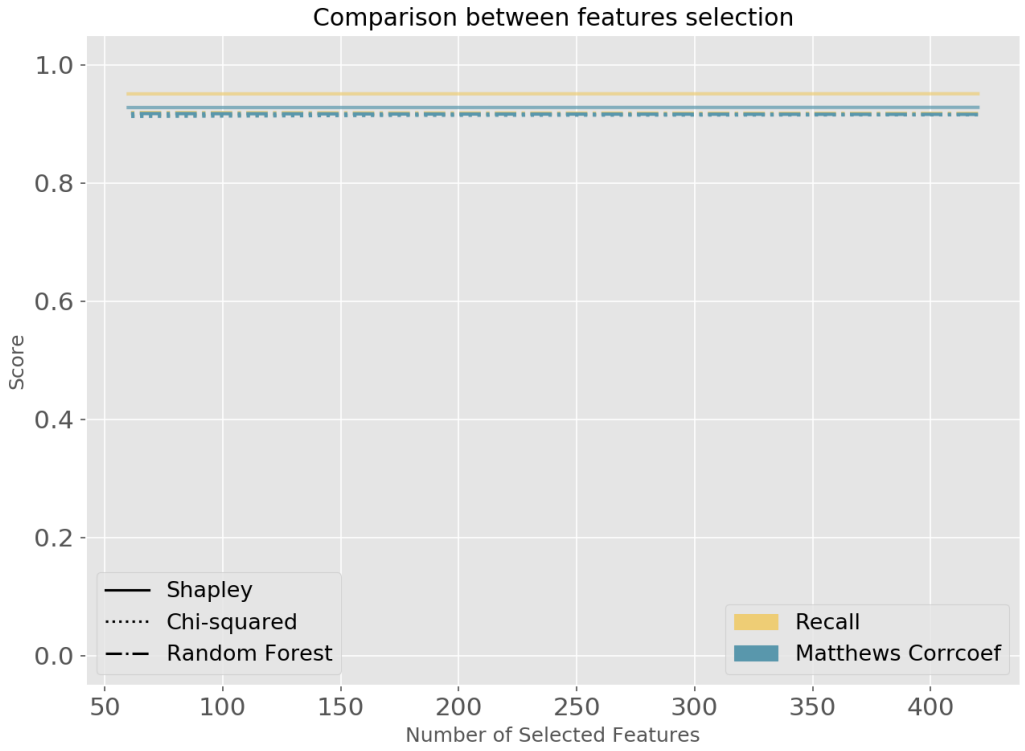

Figura 9. Utilização das métricas MCC e Revocação para comparação entre métodos de seleção de características. Os métodos valores de Shapley, Qui-quadrado e importância das características segundo método Random Forest; foram comparados variando o número de características no *BioLiP*.

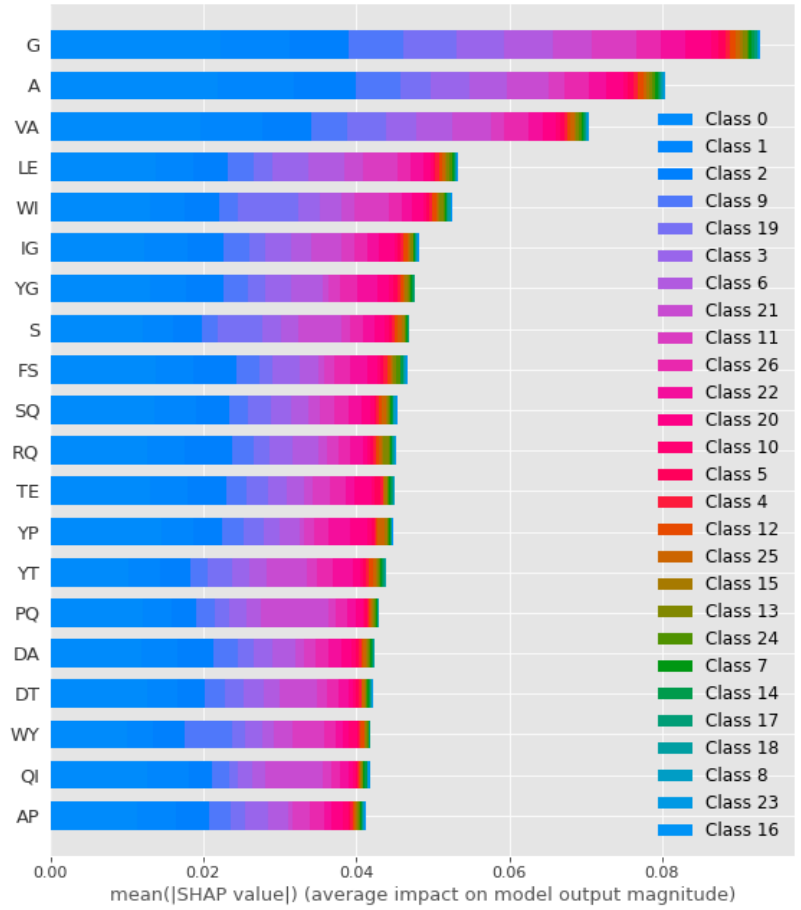

Figura 10. Figura gerada automaticamente por valor Shapley com as primeiras 20 características mais importantes para o *BioLiP*. O número de características acima de 60 não alterou os resultados, selecionamos as 60 características mais importantes, segundo *Shapley*.

3.4.1. ENGENHARIA DE CARACTERÍSTICAS NO *CATH*

Cutoff Scanning Matrix (CSM) (PIRES et al., 2011) gera vetores de características que representam padrões de distância entre resíduos de proteínas. A motivação por trás do CSM é o fato de que proteínas com diferentes dobras e funções terem distribuição de distâncias significativamente diferentes entre seus resíduos, e a similaridade de proteínas é refletida nessas distribuições de distância. O CSM gera assinaturas com base nas distâncias mínima e máxima entre carbonos alfa em uma estrutura (PIRES et al., 2011). Esta distância é escaneada com um passo incremental.

Geramos 100 conjuntos de feições usando 5Å como mínimo e 30Å como distância máxima com um passo de 2Å. Para escolher apenas um conjunto de características, realizamos validação cruzada de *10-folds* com três algoritmos diferentes, *Decision Tree,*

Random Forest e Extremely Randomized Trees com o melhor conjunto contendo 10 características no total (Anexo C).

3.4.2. ENGENHARIA E SELEÇÃO DE CARACTERÍSTICAS NO *BioLiP*

iFeature (CHEN et al., 2018) é um kit de ferramentas *Python* para calcular uma ampla gama de descritores de características estruturais e físico-químicas de proteínas e sequências peptídicas. Extraímos 420 características usando os diferentes descritores disponíveis no *iFeature*. A fim de selecionar as características mais importantes para descrever os dados do *BioLiP*, realizamos uma validação cruzada com *10-folds* e avaliamos o *MCC* variando o número de características selecionadas, com base na técnica de *grid-search* (BERGSTRA, 2012). Após cada etapa de *grid-search,* analisamos a importância das características usando o valor de Shapley (SHAPLEY, 1953). Comparamos os resultados do valor de Shapley com duas outras técnicas de seleção das caacetrísticas através do *score* classificação dos dados: Qui-quadrado e Importância da Floresta Aleatória (Figura 9). Usando essa comparação, concluímos que o valor de Shapley é bom para representar o conjunto de dados, em vez de outros métodos. Portanto, selecionamos as 60 características mais importantes, de acordo com Shapley (Figura 10 demonstra apenas 20 características por restrição de geração da imagem).

3.5. ELABORAÇÃO DOS CONJUNTOS DE TREINAMENTO E TESTE

Após a seleção das funcionalidades, foi realizado o balanceamento de classes para ambas as bases, utilizando o último nível como referência. Testamos 6 métodos de subamostragem: *Random under-sampling for the majority class, Near Miss, Condensed Nearest Neighbor Rule*, *Tomek Links, Edited Nearest Neighbor Rule*, *Neighbourhood Cleaning Rule, Cluster Centroids* (HART, 1968; MANI; ZHANG, 2003; SONNENBURG et al., 2010) e 2 métodos híbridos: *SMOTE* e *SMOTE-Tomek* (sobre e subamostragem) (BATISTA et al., 2003; CHAWLA et al., 2002). O melhor desempenho foi observado com o método *Near Miss*, que é baseado no algoritmo *nearest neighboring*. A regra heurística Near Miss seleciona as amostras da classe majoritária que possuem a menor distância média das amostras mais distantes da classe negativa. Através da biblioteca *Python imbalanced-learn* (GUILLAUME LEMAAND FERNANDO NOGUEIRA; ARIDAS, 2017), foi possível fazer um conjunto de semi-balanceamento contendo mais de 1.000 amostras no *BioLiP* e 500 amostras no *CATH*. Figura 11 mostra as distribuições das classes antes e depois das abordagens de balanceamento. Para cada nível nas abordagens locais e para o último nível nas

abordagens globais, usamos apenas classes que tivessem pelo menos 10 amostras. Isso foi empregado para garantir que uma amostra por *fold* estaria disponível para fins de validação cruzada. Portanto, o número de amostras e classes pode variar dependendo da abordagem (Anexo B e C).

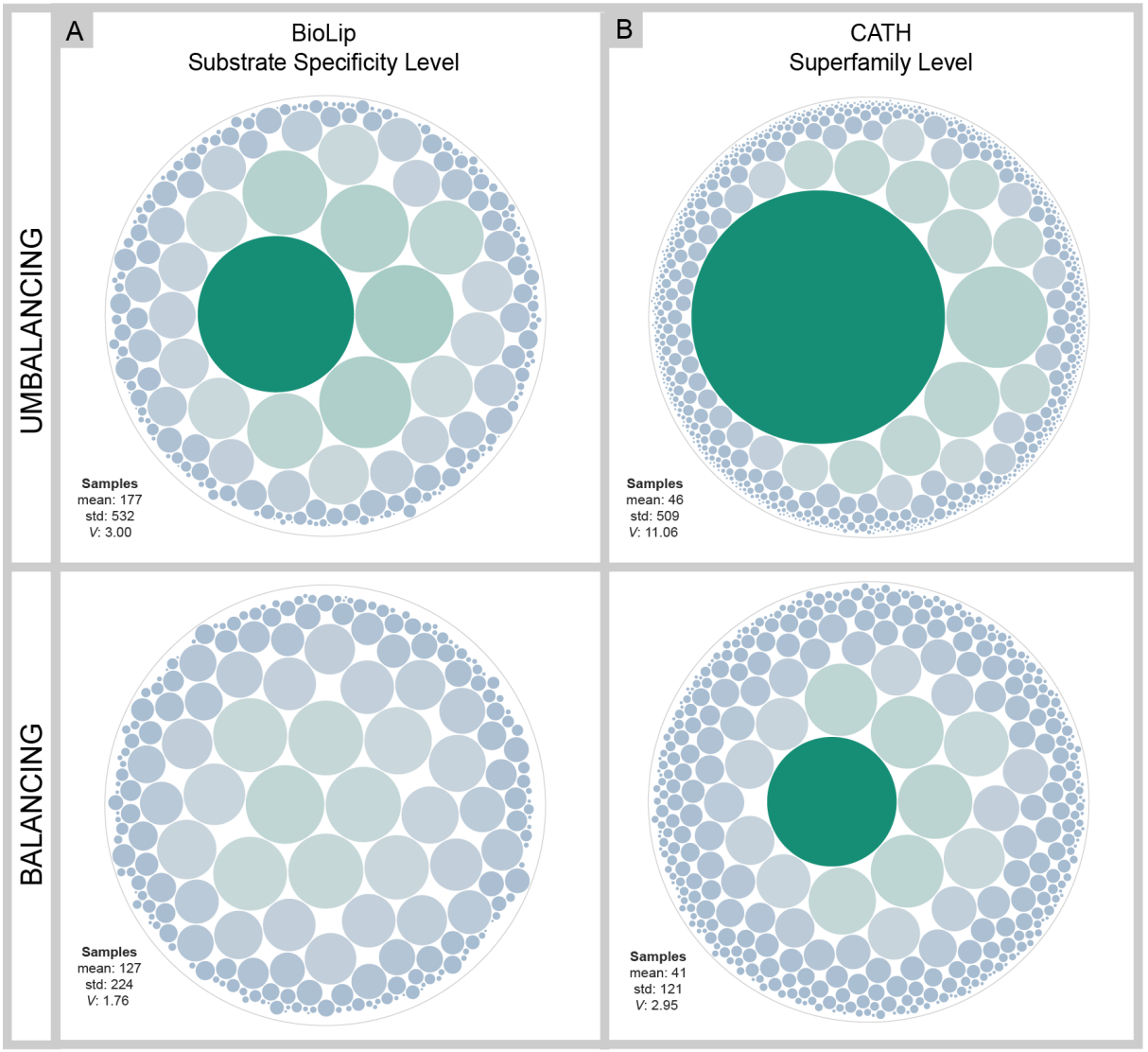

Figura 11. Estatísticas das amostras antes e após o balanceamento. O coeficiente de variação (*V*) representa a razão entre o desvio padrão (STD) e a média.

3.6. ANÁLISE DE DESEMPENHO DE ABORDAGENS HIERÁRQUICAS

Analisamos três abordagens hierárquicas: Global, Nível e Nó. Nesse sentido, os conjuntos de dados foram testados e avaliados de acordo com cada abordagem. Inicialmente, a seleção do algoritmo foi realizada com seis classificadores de aprendizado de máquina (Tabela 2). Três classificadores foram selecionados e comparados: *Decision Tree*, *Random Forest* e *Extremely Randomized Trees*.

Depois, as avaliações foram realizadas comparando as três abordagens sob validação cruzada com *10-folds (ALLEN, 1974; STONE, 1977)* e calculando a acurácia balanceada (BL ACC), MCC, área abaixo da curva ROC (AUC), *F-score* hierárquico (KIRITCHENKO et al., 2005). Como os métodos hierárquicos requerem medidas específicas para avaliar os resultados, empregamos *F-score*, Precisão e Revocação hierárquica, originalmente propostos por Kiritchenko S, Matwin et al (KIRITCHENKO et al., 2005) e recomendados por Silla e Freitas (KIRITCHENKO et al., 2005; SILLA; FREITAS, 2011). Essas medidas consideram não apenas a previsão da folha, mas também todos os ancestrais da classe em um grafo hierárquico, exceto a raiz. As equações 1 e 2 descrevem *a* precisão hierárquica *(*hP*)* e *a hierarchical recall (*hR*)*. Essas medidas combinadas são apresentadas em *F-score* hierárquico *(*hF*)* (Equação 3), em que Ci e Zi correspondem, respectivamente, a um conjunto de classes de teste e preditas para uma instância i.

$$hP = \frac{\Sigma i \, |Zi \cap Ci|}{\Sigma e \, |Zi|} \quad (1)$$

$$hR = \frac{\Sigma i \, |Zi \cap Ci|}{\Sigma e \, |Ci|} \quad (2)$$

$$hF = \frac{2 * hP * hR}{hP + hR} \quad (3)$$

4. RESULTADOS E DISCUSSÕES

Nesta seção, avaliamos as abordagens Global e Local em conjuntos de dados hierárquicos. Os modelos de aprendizado de máquina foram construídos seguindo a extração e seleção de características usando três algoritmos diferentes (*Decision Tree*, *Random Forest* e *Extremely Randomized Trees*) e avaliados sob validação cruzada de *10-folds*.

As tarefas de balanceamento e filtragem aplicadas em ambos os conjuntos de dados diminuíram a dispersão das amostras no último nível, considerando a distribuição de classes por nível e por nó. Em ambos os conjuntos de dados, obtivemos uma diminuição substancial na relação entre a média e o desvio padrão das amostras, representado pelo *V* no último nível (Figura 12).

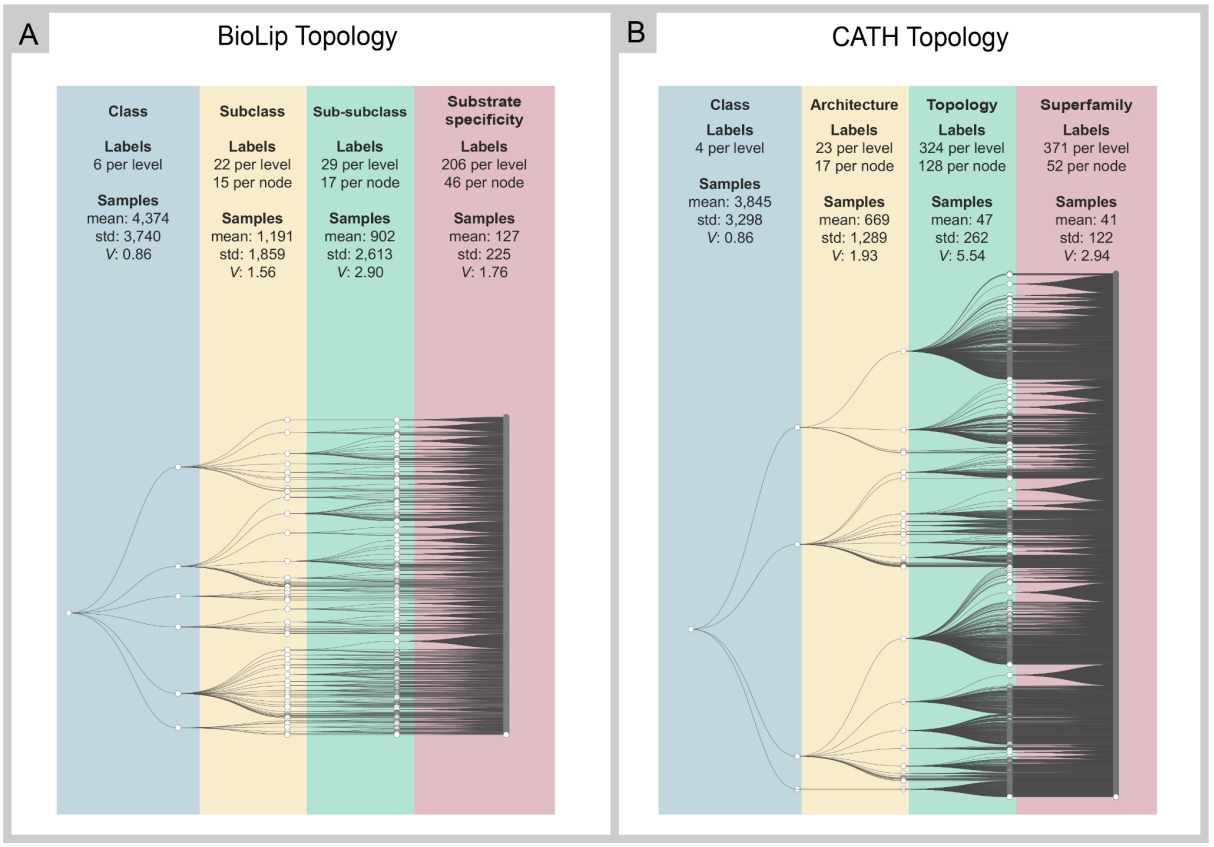

Figura 12. Topologia do *BioLiP* (A) e *CATH* (B) após balanceamento e filtragem das classes. Cada nível mostra o número de classes (rótulos) e as estatísticas resumidas de amostras por classe.

4.1. ABORDAGENS LOCAIS: COMPORTAMENTO NÍVEL A NÍVEL

Nesta seção, analisamos os resultados entre as abordagens Locais, que são divididas em Nó e Nível. Essas abordagens foram comparadas usando métricas de acurácia balanceada e MCC (Figura 13), tempo e memória usados para treinar os respectivos modelos em cada

nível (Figura 14). Métricas de desempenho convencionais, AUC e *F-score* também foram calculadas (Figura 15). Nas próximas subseções, descrevemos e analisamos o comportamento do modelo em cada nível.

Metric		BALANCED ACCURACY				MCC			
Database		BioLip		CATH		BioLip		CATH	
Approach		Node	Level	Node	Level	Node	Level	Node	Level
Level 1	DT	0.89	0.89	0.48	0.48	0.91	0.91	0.44	0.44
	RF	0.94	0.94	0.48	0.48	0.96	0.96	0.51	0.51
	ET	0.94	0.94	0.50	0.50	0.96	0.96	0.53	0.53
Level 2	DT	0.84	0.72	0.39	0.26	0.91	0.85	0.54	0.35
	RF	0.90	0.88	0.44	0.27	0.96	0.96	0.64	0.46
	ET	0.90	0.87	0.42	0.27	0.96	0.95	0.64	0.47
Level 3	DT	0.78	0.56	0.26	0.11	0.83	0.75	0.60	0.30
	RF	0.82	0.82	0.30	0.15	0.90	0.92	0.69	0.41
	ET	0.83	0.72	0.29	0.15	0.90	0.88	0.69	0.42
Level 4	DT	0.50	0.15	0.34	0.10	0.68	0.33	0.42	0.15
	RF	0.88	0.92	0.43	0.25	0.90	0.93	0.54	0.30
	ET	0.88	0.71	0.41	0.24	0.90	0.82	0.53	0.30

Figura 13. Resultados das métricas e seleção de modelos por nível para abordagens locais. Comparação de abordagens na seleção de modelos entre *Decision Tree* (*DT*), *Random Forest* (RF) e *Extra Trees* (ET) usando acurácia balanceada e MCC. As cores gradientes indicam resultados máximos em verde escuro e mínimos em branco.

Metric		Time				Memory			
Database		BioLip		CATH		BioLip		CATH	
Approach		Node	Level	Node	Level	Node	Level	Node	Level
Level 1	DT	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RF	0.54	0.41	0.22	0.32	0.93	0.99	0.66	0.18
	ET	0.33	0.22	0.09	0.13	0.94	1.00	0.65	0.18
Level 2	DT	0.04	0.00	0.00	0.00	0.17	0.00	0.00	0.00
	RF	0.49	0.50	0.30	0.35	1.00	0.99	0.38	0.18
	ET	0.24	0.35	0.19	0.15	0.90	1.00	0.51	0.18
Level 3	DT	0.00	0.03	0.00	0.01	0.08	0.00	0.01	0.00
	RF	0.59	0.60	0.64	0.63	0.62	0.88	0.51	0.48
	ET	0.46	0.38	0.50	0.29	0.66	0.93	0.81	0.55
Level 4	DT	0.03	0.07	0.00	0.02	0.08	0.00	0.02	0.01
	RF	1.00	1.00	1.00	1.00	0.62	0.88	0.69	0.75
	ET	0.68	0.55	0.87	0.48	0.66	0.93	1.00	1.00

0 1

Figura 14. Resultado da medição de tempo e memória da seleção de modelo. Comparação na seleção de modelos entre DT, RF e ET. As cores gradientes indicam resultados máximos em verde escuro e mínimos em branco.
[a]Tempo em minutos.
[b] Memória em GB.

Metric		AUC				F-SCORE			
Database		BioLip		CATH		BioLip		CATH	
Approach		Node	Level	Node	Level	Node	Level	Node	Level
Level 1	DT	0.95	0.95	0.76	0.76	0.94	0.94	0.66	0.66
	RF	0.98	0.98	0.80	0.80	0.97	0.97	0.71	0.71
	ET	0.98	0.98	0.81	0.81	0.97	0.97	0.72	0.72
Level 2	DT	0.96	0.93	0.80	0.72	0.94	0.87	0.63	0.48
	RF	0.98	0.97	0.84	0.77	0.97	0.96	0.72	0.57
	ET	0.98	0.97	0.84	0.77	0.97	0.96	0.71	0.59
Level 3	DT	0.93	0.91	0.82	0.68	0.88	0.83	0.64	0.37
	RF	0.96	0.97	0.86	0.74	0.93	0.95	0.73	0.50
	ET	0.96	0.95	0.86	0.74	0.93	0.92	0.73	0.50
Level 4	DT	0.93	0.91	0.71	0.58	0.69	0.29	0.44	0.18
	RF	0.96	0.97	0.77	0.65	0.91	0.93	0.55	0.33
	ET	0.96	0.95	0.76	0.65	0.91	0.82	0.54	0.33

0 1

Figura 15. Métricas de resultados de seleção de modelos por nível para abordagens locais. Comparação de abordagens na seleção de modelos entre DT, RF e ET usando métricas de AUC e F-score. Os valores utilizados referem-se ao último fold da validação cruzada de 10-folds. As cores do mapa de calor indicam resultados máximos em verde escuro, mínimos em tons de vermelho e ponto médio em cinza.

4.1.1. NÍVEL 1

O esforço para classificar conjuntos de dados hierárquicos no primeiro nível foi o mesmo nas duas abordagens (Nível e Nó), pois ambas geraram apenas um modelo nesse nível. A diferença entre eles está relacionada ao número de classes previstas. No conjunto de dados BioLiP, o modelo previu 6 classes; enquanto no conjunto de dados *CATH*, 4 classes foram previstas (Figura 12). A representação das classes é desbalanceada, e o número de rótulos para cada classe varia de 1 a 21 no *CATH,* e de 6 a 22 no *BioLiP* (Anexo D). Quanto às amostras, o *CATH* possui uma grande classe majoritária que representa 52,54% do total, seguida por duas classes de tamanho intermediário (24,43% e 22,86% respectivamente), e uma classe menor com 0,17% das amostras. Por sua vez, o *BioLiP* apresenta uma distribuição mais equilibrada de classes, com as três maiores abrangendo 34,11%, 32,2% e 19,91% das amostras, respectivamente. As outras três classes do *BioLiP* possuem 8,33%, 2,87%e 2,55% das amostras (Anexo D).

Não foi observada diferença (*p-valor* > 0,23, *teste T de Student*) no desempenho preditivo entre os diferentes modelos avaliados dentro de cada conjunto de dados, com *BioLiP* apresentando resultados ligeiramente melhores do que *CATH* (Figura 13). No entanto, o tempo de execução necessário para treinar *BioLiP* com *Random Forest*, foi 63% maior do que com *Extra Trees*, que apresentou desempenho semelhante, baseado em acurácia balanceada e AUC (Figura 13).

4.1.2. NÍVEL 2

No segundo nível de hierarquia, as características dos dois conjuntos de dados permaneceram semelhantes: ambos produziram um modelo com 22 e 23 classes na abordagem por nível, respectivamente. Na abordagem por nó, o *BioLiP* teve 15 classes por nó e o *CATH* teve 17 (Figura 11). *V* também está próximo de ambos os conjuntos de dados; no entanto, neste nível, excede 1, o que significa que os dados têm uma alta variância. Observamos uma diferença entre as pontuações de desempenho das abordagens por nível e nó para ambos os conjuntos de dados (Figura 13), com a abordagem por nó superando a abordagem por nível em ambos (Figura 13). Essa diferença foi observada apenas para o *CATH*.

4.1.3. NÍVEL 3

No terceiro nível, diferenças (*p-valor* < 0,01, *teste T de Student*) nas principais características dos conjuntos de dados podem ser responsáveis pelos diferentes comportamentos observados nas tarefas de seleção de modelos. Neste nível, foi evidente a elevada dispersão das classes em *CATH*, saltando de 23 classes por nível para 324, e de 17 por nó para 128. Há também uma alta variância nas amostras, com *V* aumentando quase 3 vezes em comparação com o segundo nível. As alterações no *BioLiP* foram menos significativas. Tinha 22 classes por nível no segundo nível e aumentou para 29 no terceiro. Por outro lado, havia 17 classes por nó no segundo nível, e isso o número permaneceu o mesmo no terceiro nível.

CATH apresentou diferenças (*p-valor* < 0,01, *teste T de Student*) entre as abordagens, com melhor desempenho para a abordagem por nó, apesar do maior tempo de treinamento. Nenhuma diferença foi observada para *BioLiP*. *Extra Trees* foi o mais eficiente em termos de tempo de execução do que *Random Forest* para a abordagem por nó, alcançando pontuações preditivas semelhantes, com uma pequena diferença no uso de memória. *Random Forest*

apresentou bons resultados com a abordagem por nível, incluindo menos uso de memória do que *Extra Trees*, e gastando menos tempo em comparação com a abordagem por nó. Nesse nível, ambas as abordagens usaram memória de forma semelhante, uma média de 5 GB para a abordagem Node e 6 GB para Level (Figura 14).

4.1.4. NÍVEL 4

Por fim, no último nível, por utilizarmos a técnica de semi-balanceamento por subamostragem, houve melhora em relação aos *V* das amostras em relação ao terceiro nível. CATH apresentou maior consistência em *V* comparado com *BioLiP*. Nesse nível, observou-se o mesmo padrão, ou seja, apenas *CATH* apresentou diferenças entre as abordagens (Tabela 4).

Tabela 4. *Teste T de Student* para avaliar a diferença das métricas de acurácia balanceada e MCC entre as abordagens por nó e por nível.

			p-value
BL ACC	***BioLiP***	**Level 2**	0.2278
	BioLiP	**Level 3**	0.2188
	BioLiP	**Level 4**	0.2860
BL ACC	***CATH***	**Level 2**	0.0049
	CATH	**Level 3**	0.0009
	CATH	**Level 4**	0.0100
MCC	***BioLiP***	**Level 2**	0.3788
	BioLiP	**Level 3**	0.4567
	BioLiP	**Level 4**	0.3526
MCC	***CATH***	**Level 2**	0.0019
	CATH	**Level 3**	0.0008
	CATH	**Level 4**	0.0031

Ao comparar as abordagens de nível e nó, houve uma diferença significativa no uso de memória no *CATH*. Isso pode estar associado ao número de classes ser muito maior na abordagem local por nível (371) do que na abordagem local por nó (52) (Anexo D e Tabela 5).

Tabela 5. *Teste T de Student* para avaliar a diferença entre as abordagens por nó e por nível, em relação ao tempo de treinamento e memória.

		p-value
Time	***BioLiP***	0.1751
Time	***CATH***	0.3657
Memory	***BioLiP***	0.1081
Memory	***CATH***	0.0256

Em suma, quando comparamos Abordagens Locais não houve diferenças (*p-valor* > 0,28, *teste T de Student*) (Tabela 5) entre nó e nível para *BioLiP*. Para um banco de dados que possui menos desequilíbrio entre classe e rotulagem de profundidade parcial, a abordagem de nível pode ser uma opção melhor, considerando a complexidade de implementação da abordagem por nó. No *CATH*, um banco de dados com profundidade de predição completa e *V* alto em alguns níveis, observou-se diferença (*p-valor*< 0,01, *teste T de Student*) entre as abordagens por nível e por nó. A abordagem por nó produziu modelos mais específicos, sendo a melhor opção neste caso. Na próxima seção, comparamos as abordagens locais com a mais simples de classificar hierarquias: a abordagem global.

4.2. ABO0RDAGENS GLOBAIS VERSUS LOCAIS

Para comparar de forma justa as abordagens global e local, usamos *hF* considerando o resultado do último nível das abordagens locais. A Figura 16 mostra a *hF* sob validação cruzada de *10-folds*. Em geral, como esperado, *Extra Trees* e *Random Forest* tiveram melhor desempenho em ambas as bases em todas as abordagens. Por outro lado, observamos um comportamento diferente entre os conjuntos de dados em relação às abordagens. Enquanto *CATH* teve melhor desempenho usando a abordagem global, *BioLiP* tem um melhor resultado usando abordagens locais, como confirmado por análises anteriores. Além disso, não houve diferença (*p-valor* > 0,09, *teste T de Student*) entre as abordagens global e local para *BioLiP*, enquanto há diferença para *CATH* (Tabela 6).

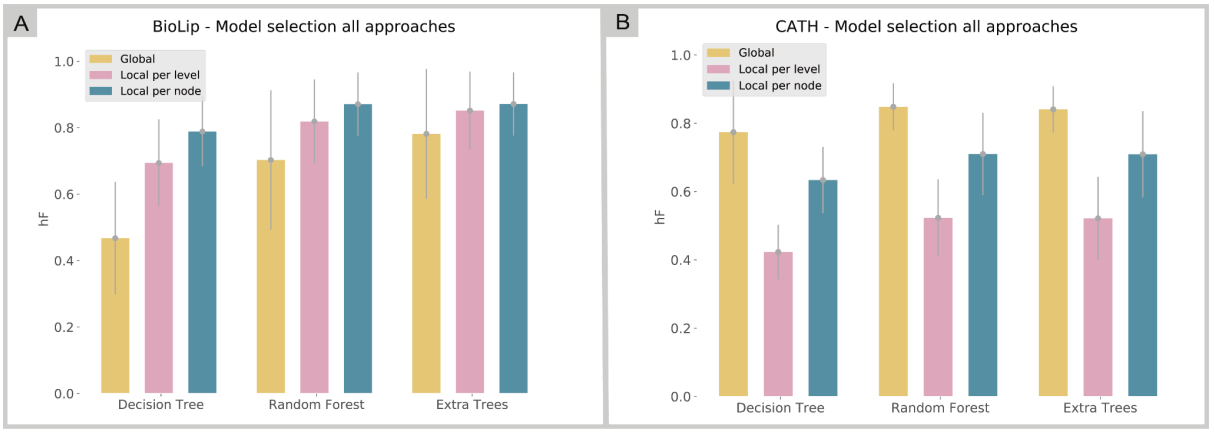

Figura 16. Comparação entre os algoritmos de *Decision Trees*, *Random Forest* e *Extra Trees* na seleção de modelos utilizando a hF. (A) Seleção de modelo realizada no conjunto de dados *BioLiP* comparando as abordagens global, local por nível e local por nó. (B) Seleção de modelo realizada no conjunto de dados *CATH* comparando as abordagens global, local por nível e local por nó. As barras de erro referem-se ao desvio padrão de desempenho para cada algoritmo.

Tabela 6. *Teste T de Student* para avaliar a diferença entre as abordagens global e locais por nó e nível.

			p-value
hF	***BioLiP***	**Nó - Global**	0.1046
		Nível - Global	0.0977
	CATH	**Nó - Global**	0.0004
		Nível - Global	0.0009

Esses bons resultados usando a abordagem global para *CATH* podem estar relacionados à rotulagem de profundidade total. Na abordagem global, podemos considerar apenas as amostras classificadas até o último nível. Outro fator que pode ter contribuído para esse resultado observado é a falta de fluxo de informações nos *CATH*. A topologia *CATH* não segue um caminho evolutivo, como outros conjuntos de dados biológicos (por exemplo, *Pfam* (BATEMAN et al., 2000), bancos de dados taxonômicos (COLE et al., 2014; DESANTIS et al., 2006; FEDERHEN, 2012; HINCHLIFF et al., 2015; PRUESSE et al., 2007), etc.); portanto, o uso de abordagens locais não são eficientes neste contexto, pois a hierarquia não reflete necessariamente as relações evolutivas de um nível para outro no mesmo ramo, nem para nós no mesmo nível. Como o 4º nível do *CATH* é diferente dos anteriores (e não possui critérios rígidos de classificação), também realizamos uma avaliação utilizando apenas os 3

primeiros níveis da hierarquia (Figura 17). Curiosamente, os resultados obtidos no nível 3 foram consistentes com aqueles usando a hierarquia completa.

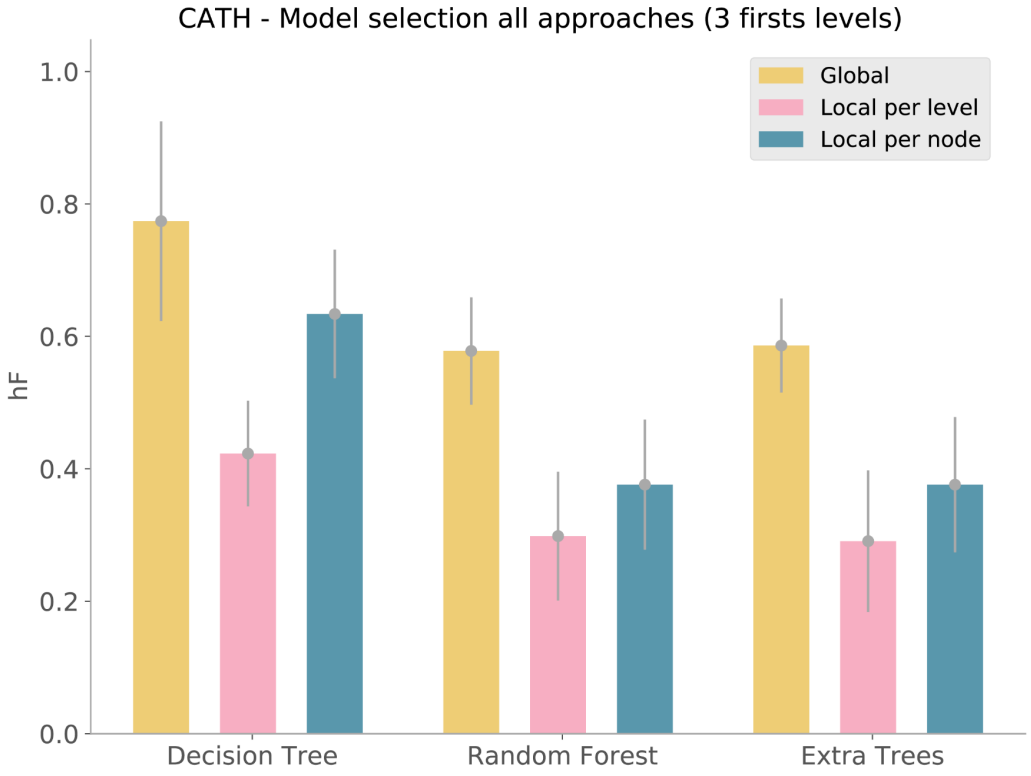

Figura 17. Comparação entre os algoritmos de *Decision Trees*, *Random Forest* e *Extra Trees* na seleção de modelos utilizando a métrica hF, usando os 3 primeiros níveis. A seleção do modelo realizada, usou os 3 primeiros níveis no conjunto de dados *CATH* comparando as abordagens global, local por nível e local por nó. As barras de erro referem-se ao desvio padrão do tempo e memória para cada algoritmo.

Para o *BioLiP*, como esperado, um número consideravelmente maior de classes na abordagem global levou a um pior desempenho em comparação aos modelos locais (Tabela 7). Além disso, as abordagens locais são mais específicas. Isso permite que vários modelos que podem manipular mais classes em comparação com a abordagem Global também lidem com classes de profundidade parcial.

Tabela 7. Comparação entre classes usando modelos treinados a partir das abordagens globais e locais.

	Global		**Level (last level)**		**Node**	
	CATH	***BioLiP***	***CATH***	***BioLiP***	***CATH***	***BioLiP***
Classes	204	589	654	206	608*	672*
Samples/class	40	15	47	127	29	68
STD**	61	35	510	225	980	252

*Calculado a partir do número médio de classes multiplicado pelo número de modelos.
**Desvio padrão do número de amostras por classe.

A Figura 18 mostra a comparação de uso de memória e tempo de treinamento entre as abordagens, bem como a diferença no número de modelos para cada abordagem local em cada nível. A abordagem por Nível produziu um modelo por nível, com uso de memória intermediário e tempos de treinamento em cada nível. A abordagem global produziu um modelo reunindo todos os níveis, com uso constante de memória e tempo de treinamento independentemente do nível. À medida que o número de classes crescia e a profundidade aumentava, mais tempo e memória eram necessários para treinar os modelos.

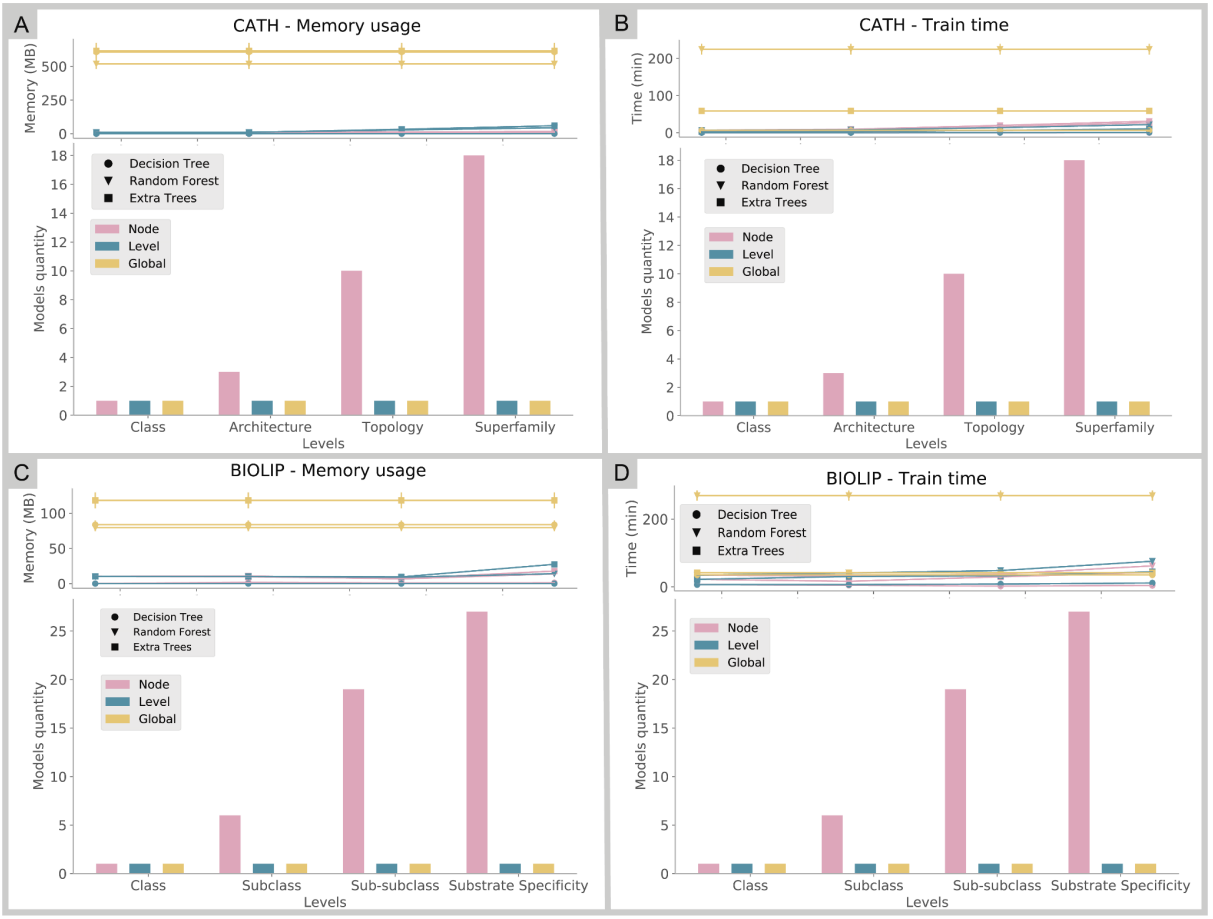

Figura 18. Número de modelos gerados pelas abordagens locais e abordagem global relacionados ao tempo e memória gastos em tarefas de treinamento. O eixo y foi dividido em duas partes: na parte superior, representamos o uso de memória, para os painéis A e C, e o tempo de treinamento, para os painéis B e D; na parte inferior, representamos as quantidades de modelos. (A) Uso de memória do *CATH*. (B) O tempo de treinamento de *CATH*. (C) Uso de memória do *BioLiP*. (D) O tempo de treinamento do *BioLiP*. As barras de erro referem-se ao *STD* da média de tempo e memória para cada algoritmo.

A abordagem por nó produziu um modelo para cada nó pai na hierarquia. Os modelos eram mais simples do que os da abordagem por nível; no entanto, foi necessário lidar com um número maior de modelos. Podemos observar que, em comparação com a abordagem global, a abordagem local consumiu menos memória (Figura 18, painéis A e C), pois tratou menos dados por modelo. No entanto, o tempo de treinamento variou drasticamente dependendo do algoritmo utilizado (Figura 18, painéis B e D). *Random Forest*, tanto para *CATH* quanto para *BioLiP*, apresentou os maiores tempos de treinamento, sendo as *Decision Tree* o algoritmo mais eficiente nesse quesito.

Como esperado, a abordagem global foi a mais exigente computacionalmente devido ao número de classes em um único modelo. Comparadas às abordagens locais por nó e nível,

a global teve 4.012 e 1.692 classes para o *CATH*, enquanto as locais por nó e nível *BioLiP* totalizaram 654 e 206 classes, respectivamente (Anexo D).

No entanto, quando a base de dados possui rotulagem completa e o objetivo da classificação está relacionado à sensibilidade, realizar os experimentos com a abordagem global pode ser uma alternativa adequada e interessante. Caso contrário, mesmo que o banco de dados tenha rotulagem de profundidade total, se o objetivo de classificação envolver especificidade, pode ser vantajoso considerar abordagens locais para obter uma classificação mais eficiente em tempo de execução.

Alternativamente, quando temos rotulagem de profundidade parcial, outros componentes devem ser considerados, incluindo a dispersão de classes e os recursos computacionais disponíveis. De acordo com nossos resultados, se o banco de dados possui baixa dispersão nos níveis ou os recursos computacionais são limitados, a abordagem de nível é mais adequada. A abordagem por nó tende a ser adequada em situações de alta dispersão de dados e quando o tempo e os recursos computacionais não são uma restrição.

4.3. DIRETRIZES PARA MODELAGEM DE CLASSIFICAÇÃO HIERÁRQUICA

Com base nos resultados discutidos acima, desenvolvemos uma diretriz inicial para auxiliar o processo decisório de modelagem de problemas de classificação hierárquica para conjuntos de dados biológicos. O fluxograma da Figura 19 descreve a escolha das abordagens a serem utilizadas, considerando os desafios de classificação detectados no conjunto de dados. Os componentes dos desafios de classificação que consideramos são a classificação do nível de profundidade, a previsão pela profundidade e as classes desequilibradas.

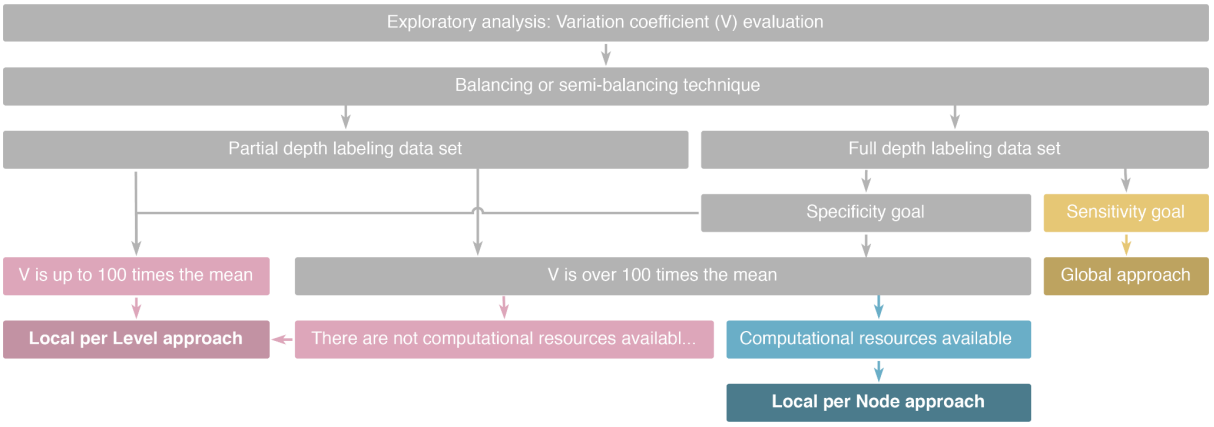

Figura 19. Diretrizes para realizar uma análise hierárquica. O fluxo de trabalho usa o Coeficiente de variação (*V*) (usado para medir a variação dos dados); os desafios geralmente enfrentados em conjuntos de dados hierárquicos, como amostras não balanceadas e previsão por profundidade

(dividida em rotulagem de profundidade parcial e rotulagem de profundidade total); e a disponibilidade de recursos computacionais para orientar a escolha de uma abordagem de classificação adequada: Global, Local por Nível ou Local por Nó.

A partir de uma análise exploratória, recomendamos analisar a natureza desbalanceada das classes no último nível, utilizando o coeficiente de variação (V), que indica a dispersão das amostras em relação à sua média. Em bancos de dados não balanceados, a abordagem ideal é aplicar uma técnica de balanceamento ou semi-balanceamento aos dados. A técnica de semi-balanceamento é preferencialmente utilizada quando as classes possuem um número limitado de amostras, para evitar a subamostragem de classes, o que dificulta a generalização do modelo.

Se o banco de dados tiver rotulagem de profundidade parcial, sugerimos a adoção de uma das abordagens Locais, usando V para orientar a decisão entre as abordagens Locais por Nível e Local por Nó. Quando V é até 100 vezes a média ($V \leq 1$) (REED; LYNN; MEADE, 2002), a abordagem Local por Nível é suficiente, alcançando bom desempenho com menos recursos computacionais. Por outro lado, amostras com $V > 1$ são consideradas altamente dispersas, consumindo mais recursos computacionais. Neste caso, aconselhamos realizar a classificação utilizando Local por Nível somente quando não houver recursos computacionais suficientes disponíveis, caso contrário sugerimos utilizar a abordagem Local por Nó.

Para bancos de dados que apresentam rotulagem de profundidade total, o critério que deve nortear os próximos passos é o objetivo da predição. A avaliação de modelos de aprendizado de máquina em termos de sensibilidade e especificidade pode ser descrita como a capacidade do preditor de detectar verdadeiros positivos e verdadeiros negativos, respectivamente. Quando o objetivo da modelagem preditiva envolve sensibilidade, a adoção de uma abordagem Global é adequada, apesar de ser mais custosa computacionalmente. Isso também é verdade quando o banco de dados tem rotulagem de profundidade parcial. Alternativamente, se o banco de dados tiver rotulagem de profundidade total e o objetivo de classificação envolver especificidade, é necessário considerar abordagens locais para obter um melhor desempenho de classificação.

Embora o objetivo dessas diretrizes gerais não seja restringir o processo de modelagem (por exemplo, uma avaliação empírica ainda é necessária), essas sugestões podem ser usadas como diretrizes iniciais para a análise de conjuntos de dados hierárquicos de até quatro níveis. É essencial começar com uma análise exploratória detalhada do conjunto de dados para identificar quais desafios de classificação hierárquica devem ser superados.

Como sugestão de uso de nossa diretriz para análises futuras, fizemos algumas recomendações com base nas características dos bancos de dados, que podem ser consistentes para bancos de dados com características semelhantes (Figuras 19 e 20). Para *CATH*, se o objetivo do trabalho for a especificidade, sugerimos que sejam priorizadas as abordagens locais. Se o objetivo principal for a sensibilidade, uma abordagem global pode ser mais adequada. Lições semelhantes podem ser potencialmente aplicadas a outros bancos de dados com a mesma estrutura e domínio, como o *Pfam* (Tabela 1). Para o *BioLiP*, sugerimos uma das abordagens locais: se houver poucos recursos computacionais disponíveis, a abordagem por nível pode ser a melhor opção; caso contrário, a abordagem por nó é uma opção interessante. Além disso, como o *KEGG* também possui desafios e estruturas semelhantes ao *BioLiP*, as mesmas lições podem ser aplicadas a ele. Observando os desafios e a estrutura dos outros bancos de dados que analisamos (Tabela 1) e aplicando nossa diretriz, sugerimos usar a abordagem local para *Silva*, *GreenGenes*, *RDP*, *OTT* e *NCBI Taxonomic*. Esperamos que a comunidade estenda essa análise para esses bancos de dados no futuro. A Figura 20 resume essas sugestões.

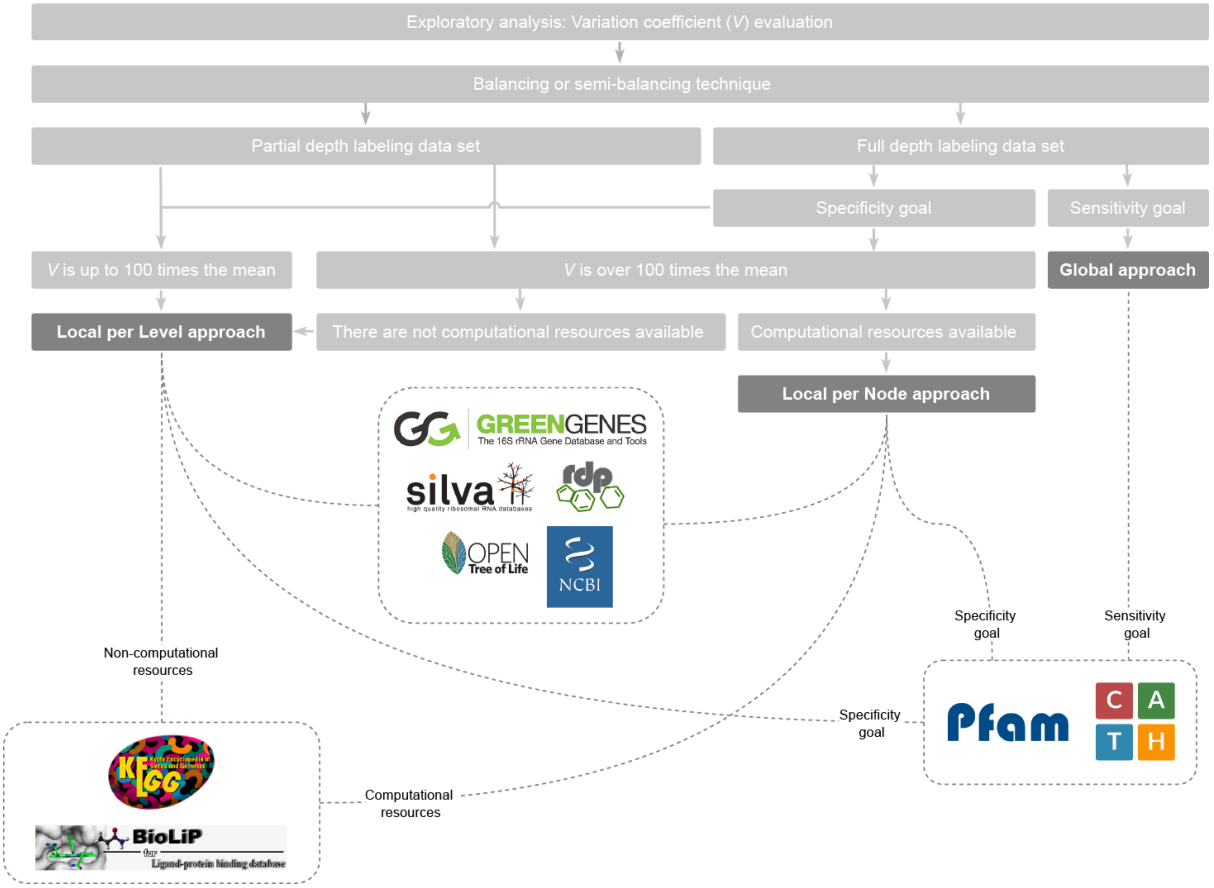

Figura 20. Resumo da extensão de abordagens de aprendizagem hierárquica para diferentes bancos de dados.

5. CONCLUSÃO

A abordagem de nível produziu um único modelo para classificar cada nível, em vez de um único grande modelo como a abordagem global. A abordagem por nó produziu um modelo para cada nó em cada nível da hierarquia, produzindo modelos mais específicos, consequentemente usando menos memória para cada modelo. Surpreendentemente, a abordagem Global apresentou melhores resultados que as abordagens locais para o banco de dados *CATH*, o que hipotetizamos poder estar vinculado a um dos componentes avaliados dos desafios hierárquicos, a previsão por profundidade.

Neste trabalho, fornecemos uma diretriz para apoiar o processo de tomada de decisão em direção a uma abordagem para obter modelos mais robustos e generalizáveis para classificar dados hierárquicos. Esta diretriz é uma proposta inicial para racionalizar a priorização da estratégia de classificação hierárquica com base nas propriedades do conjunto de dados. Esperamos fornecer evidências iniciais para apoiar mais discussões dentro da comunidade científica, o que pode levar a uma avaliação mais aprofundada em diferentes cenários biológicos. Embora este trabalho se concentre principalmente em dados biológicos, acreditamos que este guia pode ser aplicado a outros domínios do conhecimento onde dados hierárquicos estão disponíveis.

6. PERSPECTIVAS FUTURAS

Para a construção das metodologias de classificação hierárquica utilizando as três abordagens Global, Local por Nó e Local por Nível foram desenvolvidos *scripts* e bibliotecas em *python* a fim de facilitar a replicação das análises em outras bases. Em trabalhos futuros, pretendemos fornecer tais bibliotecas computacionais para auxiliar a comunidade no processo de decisão para modelar dados hierárquicos.

As diretrizes iniciais aqui propostas foram experimentadas em bases de dados de até quatro níveis. Como perspectivas futuras, as análises de bases biológicas com níveis mais profundos são sugeridas, como a base de dados do *Silva*. Além disso, pode-se também explorar os resultados em bancos de dados com características semelhantes recomendadas no trabalho (Figuras 19 e 20), aprofundando o uso das diretrizes e contribuindo com a evolução da mesma.

Para a base de dados do *Silva,* além de enfrentar outros desafios, existe também o desafio de se ter um conjunto de base de treino confiável, visto que a base de dados não é curada. Para esse trabalho vamos utilizar a técnica de aprendizagem ativa. Ela traz como princípio a possibilidade do algoritmo de classificação/regressão selecionar um subconjunto de dados que vai utilizar para treinar o preditor, a partir desse subconjunto, ele terá um melhor desempenho (SETTLES, 2009). A estratégia que será utilizada para a aprendizagem ativa será fundamentada em amostragem baseada em *pool* (LEWIS; GALE, 1994). Essa estratégia considera um grande conjunto de dados não rotulados. Serão feitas consultas selecionando um subconjunto desses dados, e será aplicado uma medida informativa, em seguida, o(s) dado(s) mais informativo(s) será(serão) selecionado(s).

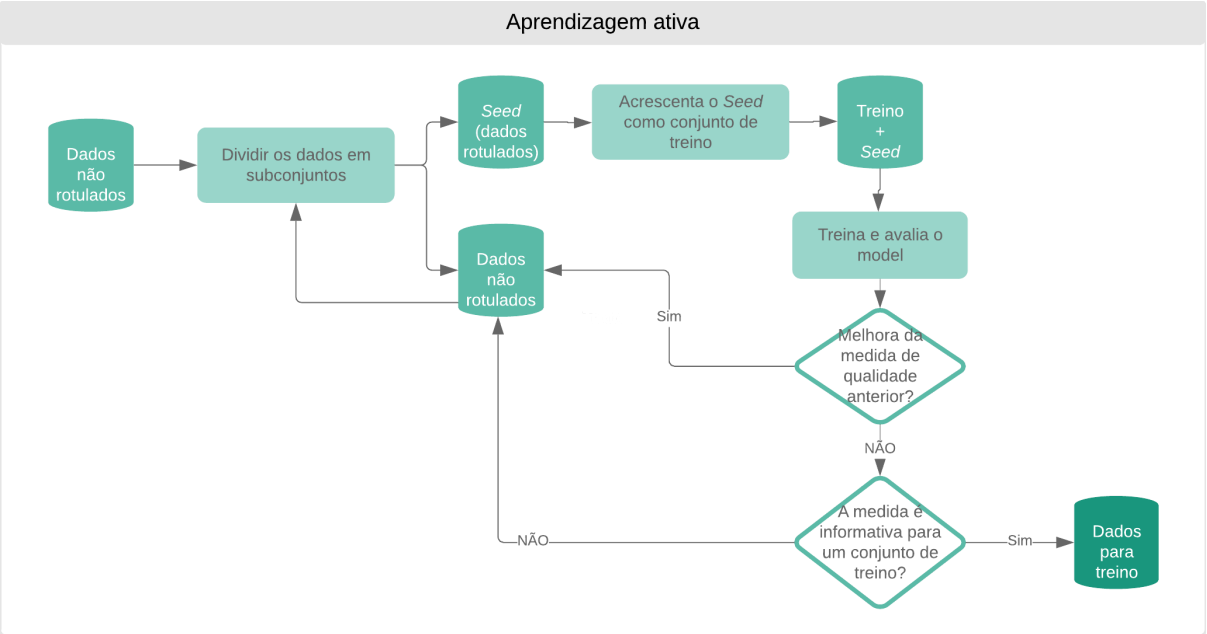

Figura 21. Algoritmo de aprendizagem ativa. Fluxograma utilizando a técnica de aprendizagem ativa, alinhada à estratégia de amostragem baseada em *pool*.

Para melhor exemplificar como serão trabalhadas cada base separadamente utilizando aprendizagem ativa, é proposto na Figura 21 um fluxograma do algoritmo. No contexto desse trabalho, todas as bases de dados inicialmente serão consideradas como conjunto de dados não rotulados. A primeira parte desse trabalho já foi realizada, onde via curadoria manual, 2000 sequências, 1% de toda a base, foram classificadas, e serão utilizadas posteriormente como *seed* inicial (DURÃES et al., 2020).

Por fim, através do melhor tipo de abordagem, é possível se concentrar em conceber as outras duas etapas propostas inicialmente no trabalho: Concepção do preditor baseado na base hierárquica e Reclassificação da base hierárquica utilizando o preditor (Figura 4). Na segunda etapa, os preditores poderão ser desenvolvidos utilizando uma ou mais abordagens de classificação hierárquica - a mais adequada segundo resultados na etapa análise de desempenho. Com isso, na terceira etapa, através dos preditores as bases de dados também poderão ser reclassificadas, tendo em vista a sua aplicabilidade e os desafios nela apresentados. Os preditores junto às novas bases reclassificadas poderão ser disponibilizados para utilização e download.

REFERÊNCIAS

AKCESME, B. Prediction of Protein Structural Classes for Low-Similarity Sequences Based On Predicted Secondary Structure. **Southeast Europe Journal of Soft Computing**, v. 4, n. 1, 1 Oct. 2015.

ALLEN, D. M. The relationship between variable selection and data agumentation and a method for prediction. **Technometrics : a journal of statistics for the physical, chemical, and engineering sciences**, v. 16, n. 1, p. 125–127, Feb. 1974.

ATTWOOD, T. K. et al. Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective. **Bioinformatics-trends and methodologies**, v. 1, p. 1–31, 2011.

BALVOČIŪTĖ, M.; HUSON, D. H. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? **BMC Genomics**, v. 18, n. Suppl 2, p. 114, 14 Mar. 2017.

BATEMAN, A. et al. The Pfam protein families database. **Nucleic Acids Research**, v. 28, n. 1, p. 263–266, 1 Jan. 2000.

BATISTA, G. E. et al. **Balancing Training Data for Automated Annotation of Keywords: a Case Study.** 2003

BEDEIAN, A. G.; MOSSHOLDER, K. W. On the use of the coefficient of variation as a measure of diversity. **Organizational Research Methods**, v. 3, n. 3, p. 285–297, Jul. 2000.

BEIKO, R. G. Microbial malaise: how can we classify the microbiome? **Trends in Microbiology**, v. 23, n. 11, p. 671–679, Nov. 2015.

BERGSTRA, J. Random Search for Hyper-Parameter Optimization. **Random Search for Hyper-Parameter Optimization**, 1 Feb. 2012.

BREIMAN, L. et al. **Classification and regression trees**. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

BREIMAN, L. Random forests. **Machine learning**, v. 45, p. 5–32, 2001.

BREITWIESER, F. P.; LU, J.; SALZBERG, S. L. A review of methods and databases for metagenomic classification and assembly. **Briefings in Bioinformatics**, v. 20, n. 4, p. 1125–1136, 19 Jul. 2019.

BROWN, C. E. Coefficient of Variation. In: **Applied multivariate statistics in geohydrology and related sciences**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. p.

155–157.

CERRI, R.; BARROS, R. C.; DE CARVALHO, A. C. P. L. F. Hierarchical multi-label classification using local neural networks. **Journal of Computer and System Sciences**, v. 80, n. 1, p. 39–56, Feb. 2014.

CERRI, R. et al. Reduction strategies for hierarchical multi-label classification in protein function prediction. **BMC Bioinformatics**, v. 17, n. 1, p. 373, 15 Sep. 2016.

CHAWLA, N. V. et al. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 1 Jun. 2002.

CHEN, Z. et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. **Bioinformatics**, v. 34, n. 14, p. 2499–2502, 15 Jul. 2018.

COLE, J. R. et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. **Nucleic Acids Research**, v. 42, n. Database issue, p. D633-42, Jan. 2014.

DAS, S. et al. CATH functional families predict functional sites in proteins. **Bioinformatics**, v. 37, n. 8, p. 1099–1106, 23 May 2021.

DAWSON, N. L. et al. CATH: an expanded resource to predict protein function through structure and sequence. **Nucleic Acids Research**, v. 45, n. D1, p. D289–D295, 4 Jan. 2017.

DA SILVA, B. M. et al. epitope3D: a machine learning method for conformational B-cell epitope prediction. **Briefings in Bioinformatics**, v. 23, n. 1, 17 Jan. 2022.

DA SILVEIRA, C. H. et al. Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. **Proteins: Structure, Function, and Bioinformatics**, v. 74, p. 727–743, 2009.

DESANTIS, T. Z. et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. **Applied and Environmental Microbiology**, v. 72, n. 7, p. 5069–5072, Jul. 2006.

DURÃES, W. Q. et al. **Taxonomic classification using active learning technique**. Congress presented at the 1st Congress of Women in Bioinformatics and Data Science Latin America. , 24 Sep. 2020. Available at: <https://drive.google.com/file/d/1Uf1e8ZqUjRkcglY1vzNRz5se6iVAk3cO/view>. Accessed: 3 Jul. 2022

FEDERHEN, S. The NCBI Taxonomy database. **Nucleic Acids Research**, v. 40, n. Database issue, p. D136-43, Jan. 2012.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine learning**, v. 63, n. 1, p. 3–42, Apr. 2006.

GUILLAUME LEMAAND FERNANDO NOGUEIRA; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. **The Journal of Machine Learning Research**, v. 18, p. 559–563, 2017.

HART, P. The condensed nearest neighbor rule (Corresp.). **IEEE Transactions on Information Theory**, v. 14, n. 3, p. 515–516, May 1968.

HENDERSON, G. et al. Improved taxonomic assignment of rumen bacterial 16S rRNA sequences using a revised SILVA taxonomic framework. **PeerJ**, v. 7, p. e6496, 5 Mar. 2019.

HINCHLIFF, C. E. et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. **Proceedings of the National Academy of Sciences of the United States of America**, v. 112, n. 41, p. 12764–12769, 13 Oct. 2015.

KANEHISA, M. et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. **Nucleic Acids Research**, v. 45, n. D1, p. D353–D361, 4 Jan. 2017.

KIRITCHENKO, S. et al. **Functional annotation of genes using hierarchical text categorization**. 2005

KOSMOPOULOS, A. et al. Evaluation measures for hierarchical classification: a unified view and novel approaches. **Data mining and knowledge discovery**, v. 29, n. 3, p. 820–865, May 2015.

KOWSARI, K. et al. **Hdltex: Hierarchical deep learning for text classification**. 2017

KULMANOV, M. et al. Semantic similarity and machine learning with ontologies. **Briefings in Bioinformatics**, v. 22, n. 4, 20 Jul. 2021.

LEWIS, D. D.; GALE, W. A. **A sequential algorithm for training text classifiers**. 1994

LUNDBERG, S.; LEE, S.-I. A Unified Approach to Interpreting Model Predictions. **arXiv**, 2017.

MANI, I.; ZHANG, I. **kNN approach to unbalanced data distributions: a case study involving information extraction**. 2003

MURZIN, A. G. et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. **Journal of Molecular Biology**, v. 247, n. 4, p. 536–540, 7 Apr. 1995.

MYUNG, Y. et al. mCSM-AB2: guiding rational antibody design using graph-based signatures. **Bioinformatics**, v. 36, n. 5, p. 1453–1459, 1 Mar. 2020.

NAKANO, F. K. et al. **Top-down strategies for hierarchical classification of transposable elements with neural networks**. 2017

OGATA, H. et al. KEGG: kyoto encyclopedia of genes and genomes. **Nucleic Acids Research**, v. 27, n. 1, p. 29–34, 1 Jan. 1999.

ORENGO, C. A. et al. CATH--a hierarchic classification of protein domain structures. **Structure**, v. 5, n. 8, p. 1093–1108, 15 Aug. 1997.

PANTA, M. et al. ClassifyTE: A stacking based prediction of hierarchical classification of transposable elements. **Bioinformatics**, 2 Mar. 2021.

PARKS, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. **Nature Biotechnology**, v. 36, n. 10, p. 996–1004, Nov. 2018.

PEARL, F. M. G. et al. The CATH database: an extended protein family resource for structural and functional genomics. **Nucleic Acids Research**, v. 31, n. 1, p. 452–455, 1 Jan. 2003.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in {P}ython. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PIRES, D. E. V.; ASCHER, D. B.; BLUNDELL, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. **Bioinformatics**, v. 30, n. 3, p. 335–342, 1 Feb. 2014a.

PIRES, D. E. V.; ASCHER, D. B.; BLUNDELL, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. **Nucleic Acids Research**, v. 42, n. Web Server issue, p. W314-9, Jul. 2014b.

PIRES, D. E. V.; ASCHER, D. B. CSM-lig: a web server for assessing and comparing protein-small molecule affinities. **Nucleic Acids Research**, v. 44, n. W1, p. W557-61, 8 Jul. 2016.

PIRES, D. E. V.; ASCHER, D. B. mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. **Nucleic Acids Research**, v. 45, n. W1, p. W241–W246, 3 Jul. 2017.

PIRES, D. E. V.; BLUNDELL, T. L.; ASCHER, D. B. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. **Scientific Reports**, v. 6, p. 29575, 7 Jul. 2016.

PIRES, D. E. V.; RODRIGUES, C. H. M.; ASCHER, D. B. mCSM-membrane: predicting the effects of mutations on transmembrane proteins. **Nucleic Acids Research**, v. 48, n. W1, p. W147–W153, 2 Jul. 2020.

PIRES, D. E. V. et al. Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. **BMC Genomics**, v. 12 Suppl 4, p. S12, 22 Dec. 2011.

PIRES, D. E. V. et al. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. **Bioinformatics**, v. 29, n. 7, p. 855–861, 1 Apr. 2013.

PRUESSE, E. et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. **Nucleic Acids Research**, v. 35, n. 21, p. 7188–7196, 18 Oct. 2007.

PYBUS, M. et al. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. **Bioinformatics**, v. 31, n. 24, p. 3946–3952, 15 Dec. 2015.

QUAST, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. **Nucleic Acids Research**, v. 41, n. Database issue, p. D590-6, Jan. 2013.

REED, G. F.; LYNN, F.; MEADE, B. D. Use of coefficient of variation in assessing variability of quantitative assays. **Clinical and Vaccine Immunology**, v. 9, p. 1235–1239, 2002.

RODRIGUES, C. H.; ASCHER, D. B.; PIRES, D. E. Kinact: a computational approach for predicting activating missense mutations in protein kinases. **Nucleic Acids Research**, v. 46, n. W1, p. W127–W132, 2 Jul. 2018.

SANDARUWAN, P. D.; WANNIGE, C. T. An improved deep learning model for hierarchical classification of protein families. **Plos One**, v. 16, n. 10, p. e0258625, 20 Oct. 2021.

SETTLES, B. Active learning literature survey. 2009.

SHAPLEY, L. S. A value for n-person games. **Contributions to the Theory of Games**, v. 2, p. 307–317, 1953.

SHEN, Y.; TANG, J.; GUO, F. Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. **Journal of Theoretical Biology**, v. 462, p. 230–239, 7 Feb. 2019.

SILLA, C. N.; FREITAS, A. A. A survey of hierarchical classification across different application domains. **Data mining and knowledge discovery**, v. 22, n. 1–2, p. 31–72, Jan. 2011.

SÖHNGEN, C. et al. Bac Dive--the bacterial diversity metadatabase in 2016. **Nucleic acids research**, v. 44, p. D581–D585, 2016.

SONG, J. et al. iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. **Briefings in Bioinformatics**, v. 20, n. 2, p. 638–658, 25 Mar. 2019.

SONNENBURG, S. et al. The SHOGUN machine learning toolbox. **The Journal of Machine Learning Research**, v. 11, p. 1799–1802, 2010.

STONE, M. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 39, n. 1, p. 44–47, Sep. 1977.

STRODTHOFF, N. et al. UDSMProt: universal deep sequence models for protein classification. **Bioinformatics**, v. 36, n. 8, p. 2401–2409, 15 Apr. 2020.

TIAN, L. Inferences on the common coefficient of variation. **Statistics in Medicine**, v. 24, n. 14, p. 2213–2220, 30 Jul. 2005.

WEI, L. et al. A novel hierarchical selective ensemble classifier with bioinformatics application. **Artificial Intelligence in Medicine**, v. 83, p. 82–90, Nov. 2017.

XIONG, D.; ZENG, J.; GONG, H. A deep learning framework for improving long-range residue-residue contact prediction using a hierarchical strategy. **Bioinformatics**, v. 33, n. 17, p. 2675–2683, 1 Sep. 2017.

YANG, J.; ROY, A.; ZHANG, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. **Nucleic Acids Research**, v. 41, n. Database issue, p.

D1096-103, Jan. 2013.

YILMAZ, P. et al. The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. **Nucleic Acids Research**, v. 42, n. Database issue, p. D643-8, Jan. 2014.

YOON, S.-H. et al. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. **International Journal of Systematic and Evolutionary Microbiology**, v. 67, n. 5, p. 1613–1617, 30 May 2017.

ZHANG, Y.; WANG, Z.; WANG, Y. Multi-hierarchical profiling: an emerging and quantitative approach to characterizing diverse biological networks. **Briefings in Bioinformatics**, v. 18, n. 1, p. 57–68, Jan. 2017.

ANEXO A - REPRESENTAÇÃO DA BASE DE DADOS SILVA POR NÍVEL

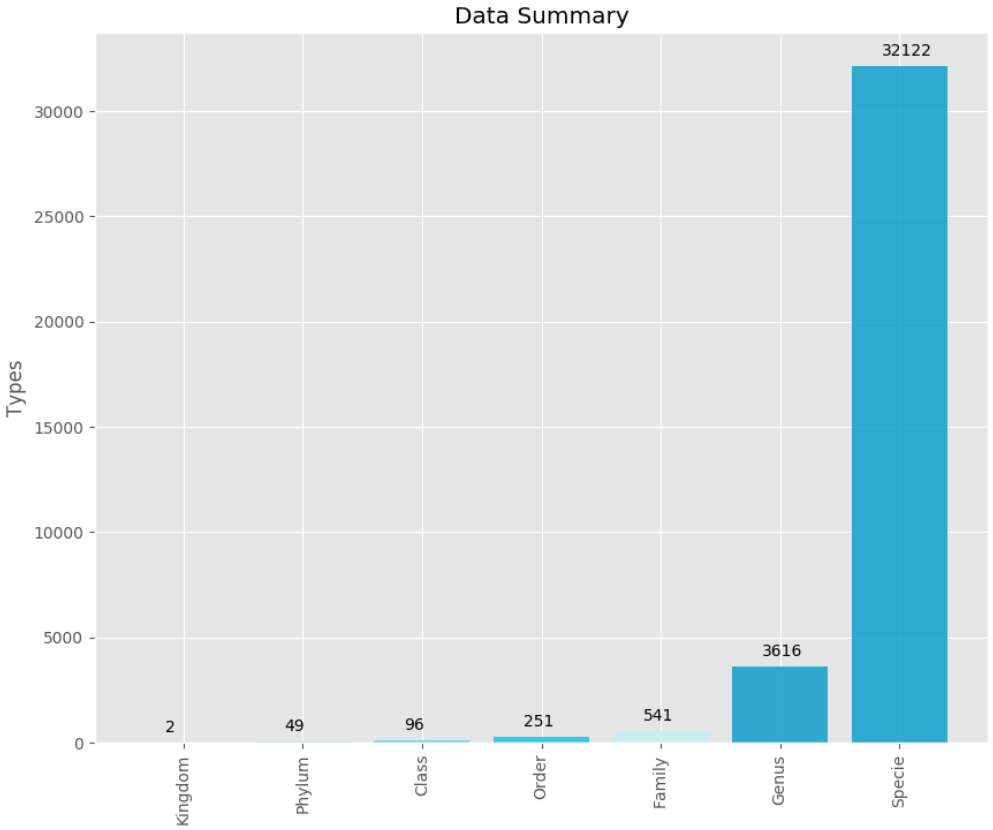

Figura A1. Resumo da quantidade de rótulos por nível taxonômico na base Silva.

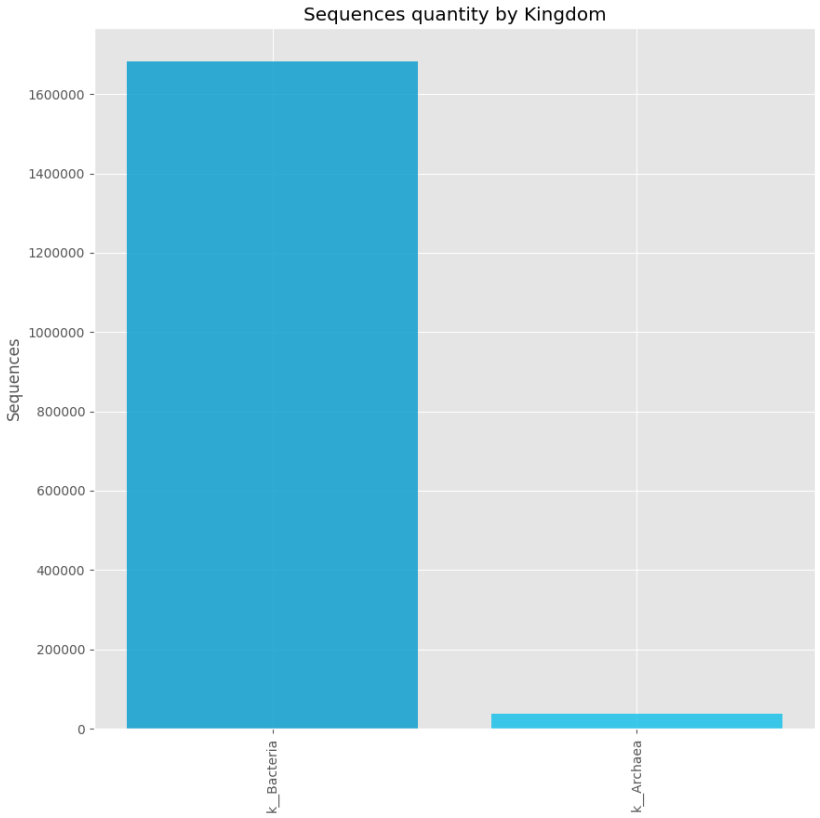

Figura A2. Classificação dos rótulos pelo nível Domínio, Archaea e Bacteria na base Silva.

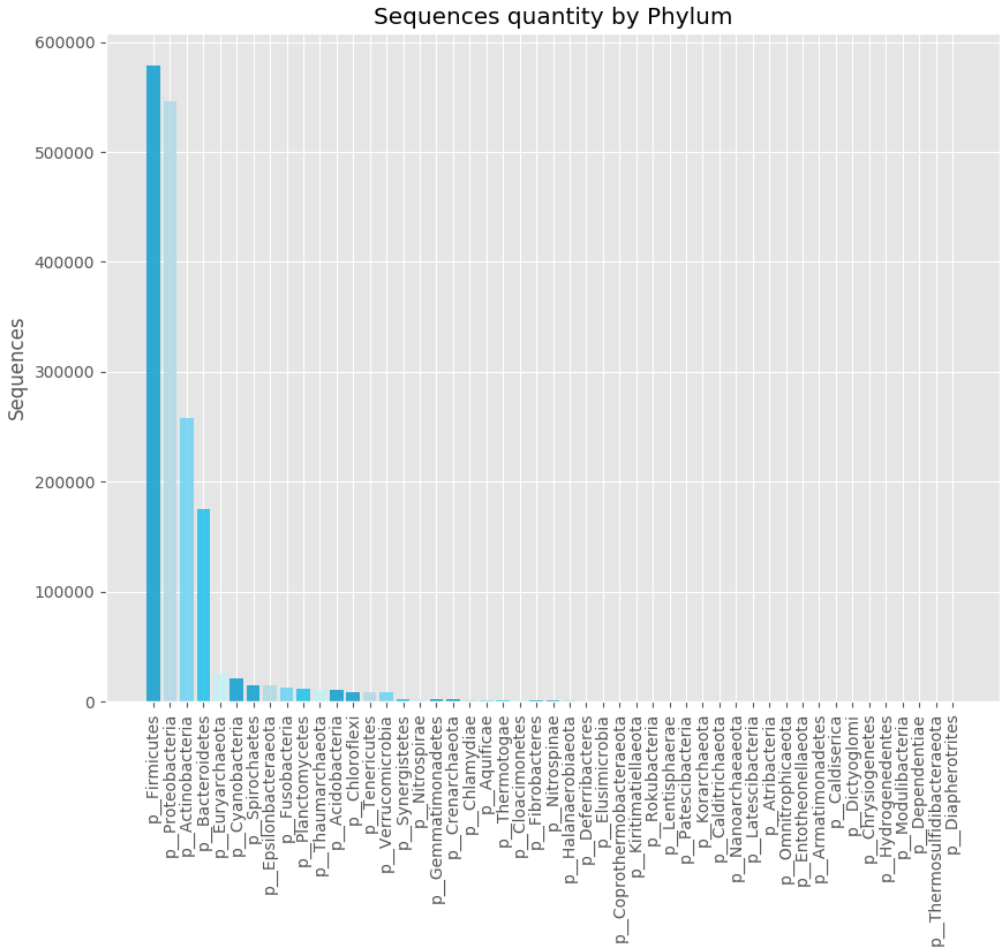

Figura A3. Quantidade de sequências classificados por rótulos no nível Filo na base Silva.

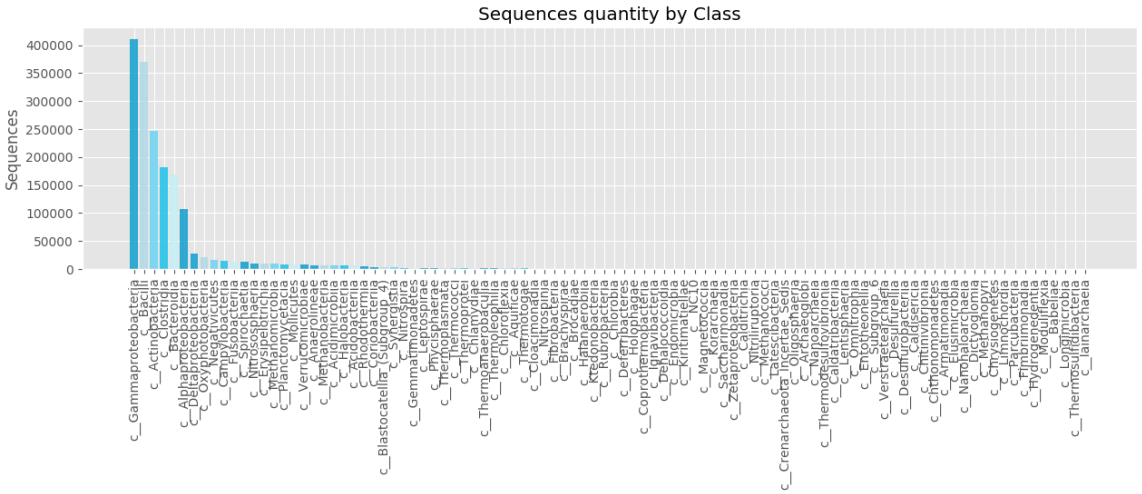

Figura A4. Quantidade de sequências classificados por rótulos no nível Classe na base Silva.

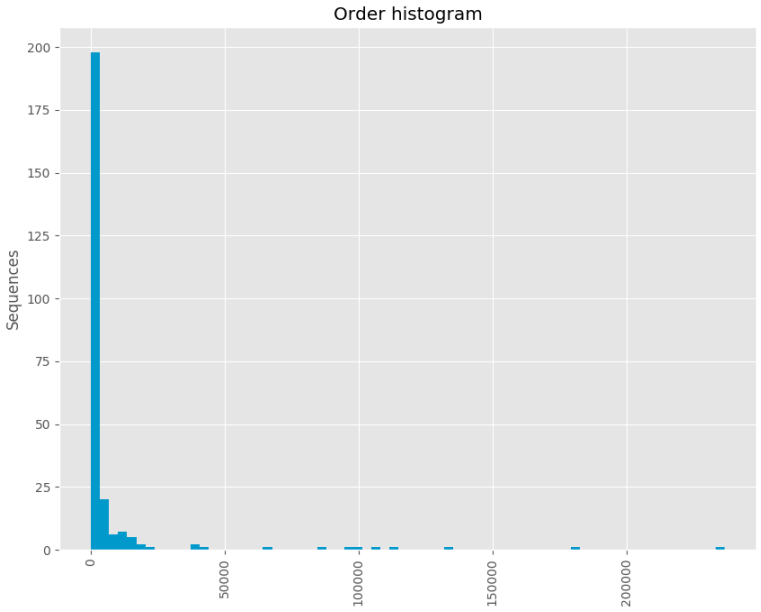

Figura A5. Distribuição de sequências por classificação a nível de Ordem na base Silva.

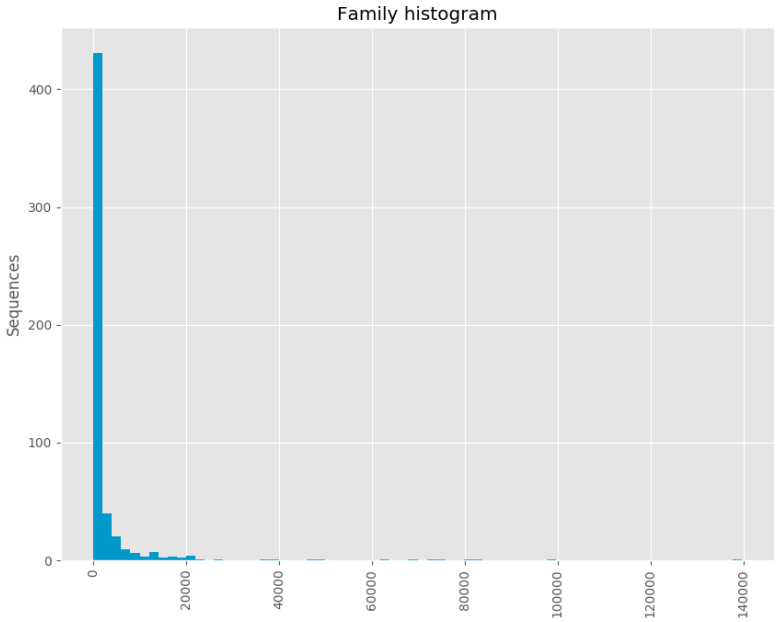

Figura A6. Distribuição de sequências por classificação a nível de Família na base Silva.

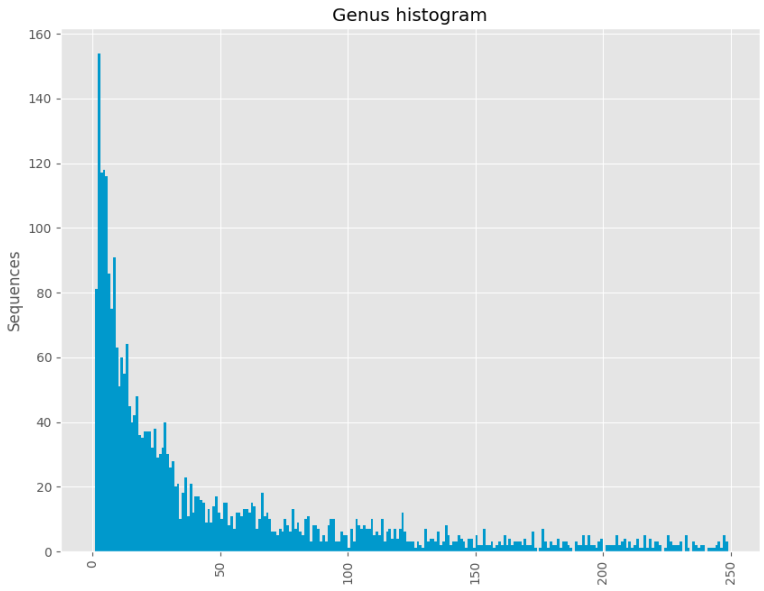

Figura A7. Distribuição de sequências por classificação a nível de Gênero na base Silva.

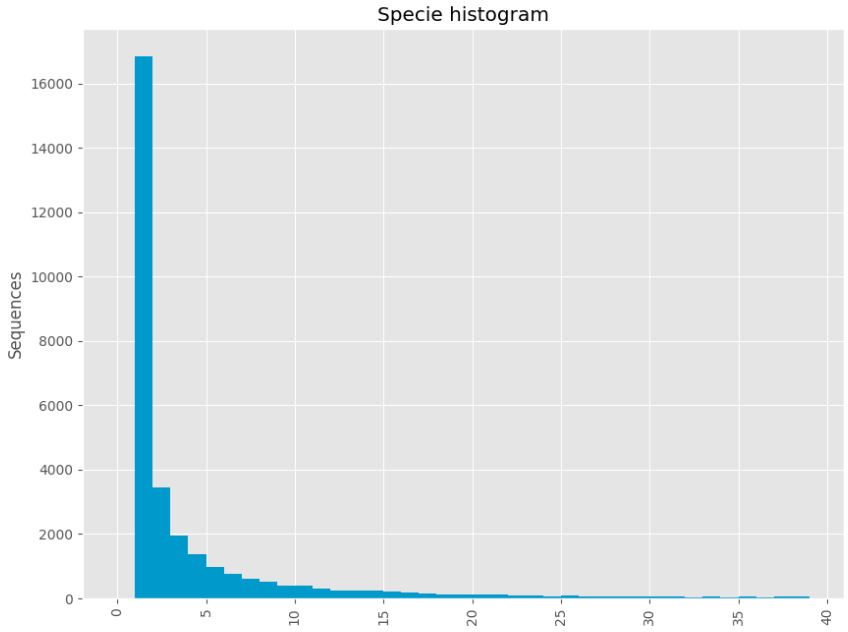

Figura A8. Distribuição de sequências por classificação a nível de Espécie na base Silva.

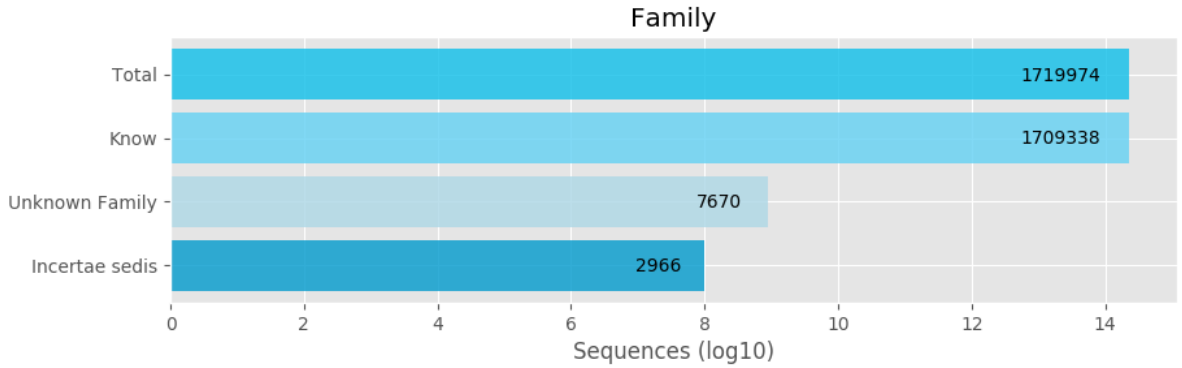

Figura A9. Distribuição de classificação de sequências por Família. Classificação das sequências seguindo famílias conhecidas (Know), desconhecidas (Unknown Family) e quando a classificação é incerta (Incertae sedis)

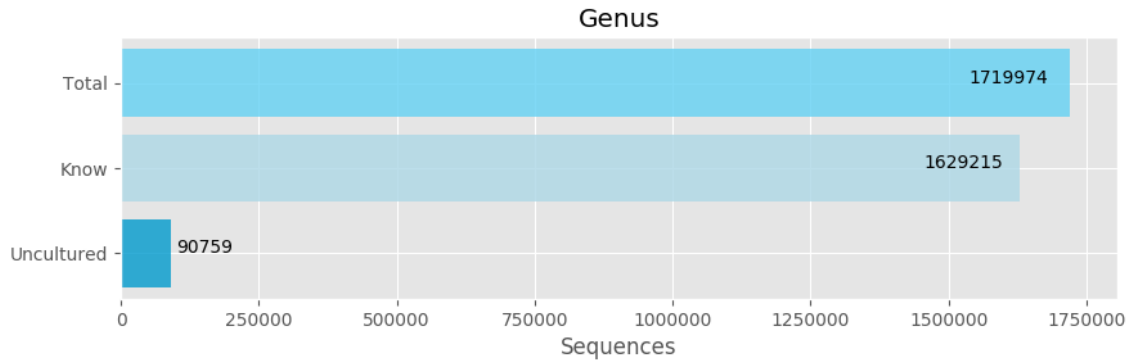

Figura A10. Distribuição de classificação de sequências por Gênero. Classificação das sequências seguindo gêneros conhecidos (Know) e desconhecidos (Uncultured)

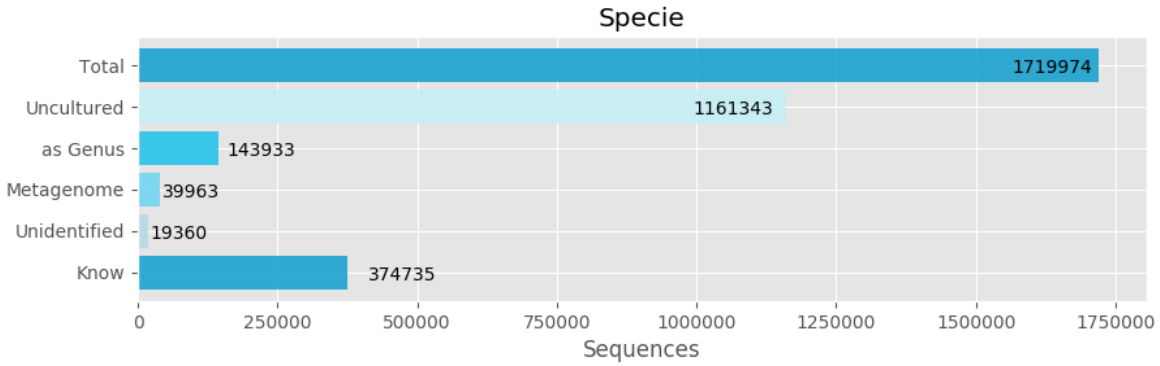

Figura A11. Distribuição de classificação de sequências por Espécies. Classificação das sequências seguindo gêneros conhecidos (Know), desconhecidos (Uncultured e Metagenome), não identificados (Unidentified), e classificados apenas pelo nível Gênero (as Genus)

ANEXO B - REPRESENTAÇÃO DAS CLASSES POR NÍVEL ANTES DO BALANCEAMENTO

		1º Level		2º Level		3º Level		4º Level	
		CATH	BIOLIP	CATH	BIOLIP	CATH	BIOLIP	CATH	BIOLIP
Classes	per level	4	6	26	23	520	32	654	206
	per node	-	-	10	11	46	6	9	23
Samples	mean	7,684.25	6,087.67	1,182.19	1,578.52	59.00	1,121.69	46.00	177.31
	std	6,622.29	4,838.16	2,310.56	2,351.38	291.00	2,868.25	509.00	532.99
	vc	0.86	0.79	1.95	1.49	4.93	2.56	11.07	3.01

Representação das classes no Nível 1				
	Class	Samples	Number of labels	Representation (%)
CATH	B	7.349	5	23.91
	C	6.758	21	21.99
	D	16.371	14	53.26
	E	259	1	0.84
BIOLIP	A	8.057	23	22.06
	B	11.026	10	30.19
	C	11.9	13	32.58
	D	2.687	8	7.36
	E	1.45	7	3.97
	F	1.406	6	3.85

Representação das classes no Nível 2				
	Class	Samples	Number of labels	Representation (%)
CATH	B	6464	361	21.03
	C	13	3	0.04
	D	3524	118	11.47
	E	502	6	1.63
	F	5780	250	18.80

	G	8332	161	27.11
	H	429	12	1.40
	I	28	6	0.09
	J	3165	51	10.30
	K	16	1	0.05
	L	208	18	0.68
	M	14	1	0.05
	N	205	3	0.67
	O	1474	163	4.80
	P	31	1	0.10
	Q	18	3	0.06
	R	1	1	0.00
	S	7	1	0.02
	T	40	1	0.13
	U	69	1	0.22
	V	137	1	0.45
	W	10	1	0.03
	X	14	1	0.05
	Y	124	3	0.40
	Z	131	26	0.43
	AB	1	1	0.00
BIOLIP	A	7354	19	20.13
	B	4424	7	12.11
	C	2935	8	8.04
	D	6425	21	17.59
	E	3034	7	8.31
	F	1823	9	4.99
	G	6326	15	17.32
	H	475	8	1.30
	I	192	3	0.53
	J	241	4	0.66
	K	422	4	1.16
	L	44	6	0.12
	M	307	4	0.84
	N	1457	13	3.99

	O	173	1	0.47
	P	96	3	0.26
	Q	188	5	0.51
	R	164	4	0.45
	S	1	1	0.00
	T	21	1	0.06
	U	3	1	0.01
	V	18	1	0.05
	W	183	1	0.50

Representação das classes no Nível 3				
	Class	Samples	Number of labels	Representation (%)
C A T H	4	16	3	0.05
	5	400	102	1.30
	6	1	1	0.00
	7	10	1	0.03
	8	293	118	0.95
	9	47	9	0.15
	10	2082	274	6.77
	11	2	1	0.01
	12	43	9	0.14
	14	3	1	0.01
	15	82	3	0.27
	16	17	2	0.06
	18	3	1	0.01
	19	6	1	0.02
	20	1618	73	5.26
	21	46	5	0.15
	22	4	1	0.01

	24	2	1	0.01
	25	222	62	0.72
	26	1	1	0.00
	27	3	1	0.01
	28	82	26	0.27
	29	172	23	0.56
	30	987	109	3.21
	31	13	7	0.04
	32	1	1	0.00
	33	26	3	0.08
	34	6	3	0.02
	35	5	1	0.02
	36	4	3	0.01
	37	25	3	0.08
	38	5	1	0.02
	39	33	1	0.11
	40	2535	382	8.25
	42	112	6	0.36
	43	24	2	0.08
	44	1	1	0.00
	45	17	1	0.06
	46	1	1	0.00
	47	87	2	0.28
	49	1	1	0.00
	50	5087	363	16.55
	51	1	1	0.00
	54	4	2	0.01
	55	3	1	0.01

	56	23	8	0.07
	58	359	191	1.17
	59	10	3	0.03
	60	150	24	0.49
	62	1	1	0.00
	63	5	1	0.02
	65	1	1	0.00
	66	2	1	0.01
	67	5	3	0.02
	69	3	1	0.01
	70	1227	278	3.99
	75	5	2	0.02
	76	13	1	0.04
	77	5	1	0.02
	78	2	1	0.01
	79	77	4	0.25
	80	68	5	0.22
	81	9	3	0.03
	82	3	1	0.01
	85	6	2	0.02
	87	1	1	0.00
	89	4	2	0.01
	90	56	5	0.18
	91	26	9	0.08
	92	1	1	0.00
	93	1	1	0.00
	95	1	1	0.00
	98	77	11	0.25

	100	90	3	0.29
	101	6	1	0.02
	105	34	3	0.11
	109	57	3	0.19
	110	285	18	0.93
	120	1260	182	4.10
	128	168	60	0.55
	129	131	10	0.43
	130	93	8	0.30
	132	33	13	0.11
	135	2	1	0.01
	140	140	18	0.46
	141	4	1	0.01
	142	6	2	0.02
	144	3	1	0.01
	148	1	1	0.00
	150	349	80	1.14
	155	14	1	0.05
	160	280	76	0.91
	166	1	1	0.00
	167	9	3	0.03
	168	4	2	0.01
	169	1	1	0.00
	170	56	6	0.18
	175	6	2	0.02
	176	16	1	0.05
	180	187	9	0.61
	182	2	1	0.01

	185	1	1	0.00
	189	3	1	0.01
	190	584	24	1.90
	195	1	1	0.00
	196	6	3	0.02
	198	3	1	0.01
	199	3	1	0.01
	200	294	21	0.96
	209	6	1	0.02
	210	46	4	0.15
	215	4	2	0.01
	220	124	14	0.40
	225	32	4	0.10
	226	105	4	0.34
	228	19	2	0.06
	230	128	11	0.42
	238	178	21	0.58
	239	2	1	0.01
	240	74	10	0.24
	241	1	1	0.00
	245	17	1	0.06
	246	40	21	0.13
	249	1	1	0.00
	250	74	2	0.24
	260	172	19	0.56
	268	6	2	0.02
	269	2	2	0.01
	270	55	5	0.18

	272	19	6	0.06
	274	22	10	0.07
	275	27	6	0.09
	280	37	3	0.12
	285	3	2	0.01
	286	15	10	0.05
	287	321	99	1.04
	288	2	2	0.01
	290	42	7	0.14
	300	133	31	0.43
	305	1	1	0.00
	309	38	1	0.12
	310	114	27	0.37
	320	36	3	0.12
	330	37	10	0.12
	340	35	8	0.11
	342	5	1	0.02
	350	41	2	0.13
	357	159	16	0.52
	360	105	11	0.34
	365	27	2	0.09
	366	18	2	0.06
	367	7	1	0.02
	370	7	2	0.02
	372	1	1	0.00
	375	9	1	0.03
	379	17	3	0.06
	380	16	2	0.05

	386	2	1	0.01
	387	1	1	0.00
	390	177	13	0.58
	395	4	1	0.01
	400	32	2	0.10
	405	10	3	0.03
	410	24	2	0.08
	412	1	1	0.00
	413	10	1	0.03
	418	31	8	0.10
	420	367	47	1.19
	428	24	5	0.08
	429	31	1	0.10
	430	40	4	0.13
	437	23	3	0.07
	439	7	2	0.02
	440	14	1	0.05
	441	2	1	0.01
	442	1	1	0.00
	443	12	3	0.04
	449	5	1	0.02
	450	488	66	1.59
	455	3	1	0.01
	457	6	5	0.02
	460	70	11	0.23
	462	19	3	0.06
	465	35	3	0.11
	468	2	1	0.01

	470	150	11	0.49
	472	70	16	0.23
	479	14	3	0.05
	480	7	2	0.02
	486	7	1	0.02
	489	2	1	0.01
	490	101	16	0.33
	497	31	1	0.10
	499	8	2	0.03
	500	32	4	0.10
	505	48	4	0.16
	506	8	2	0.03
	510	154	4	0.50
	519	2	1	0.01
	520	36	5	0.12
	525	6	1	0.02
	530	113	8	0.37
	532	4	1	0.01
	533	55	3	0.18
	540	61	3	0.20
	550	78	6	0.25
	555	17	1	0.06
	559	38	4	0.12
	560	16	1	0.05
	565	69	5	0.22
	569	3	1	0.01
	570	13	1	0.04
	572	7	1	0.02

	575	4	1	0.01
	579	5	1	0.02
	580	56	2	0.18
	590	21	4	0.07
	599	2	1	0.01
	600	67	4	0.22
	601	1	1	0.00
	605	37	1	0.12
	606	4	2	0.01
	610	1	1	0.00
	620	24	3	0.08
	630	358	15	1.16
	640	198	5	0.64
	645	5	2	0.02
	650	15	1	0.05
	660	23	4	0.07
	670	3	1	0.01
	680	2	1	0.01
	690	2	1	0.01
	700	23	5	0.07
	710	104	3	0.34
	718	15	1	0.05
	720	91	14	0.30
	730	34	3	0.11
	740	11	1	0.04
	750	45	20	0.15
	760	70	2	0.23
	770	3	1	0.01

	780	17	3	0.06
	790	6	2	0.02
	800	36	2	0.12
	810	11	1	0.04
	820	6	1	0.02
	830	58	1	0.19
	840	12	1	0.04
	850	28	2	0.09
	860	17	2	0.06
	870	35	4	0.11
	880	4	1	0.01
	890	26	10	0.08
	900	26	2	0.08
	910	35	4	0.11
	920	62	9	0.20
	930	103	10	0.34
	940	16	4	0.05
	950	42	3	0.14
	960	23	5	0.07
	970	24	5	0.08
	980	28	5	0.09
	990	10	1	0.03
	1000	23	8	0.07
	1010	30	3	0.10
	1020	9	2	0.03
	1030	14	1	0.05
	1040	69	5	0.22
	1050	147	13	0.48

	1060	21	3	0.07
	1070	49	3	0.16
	1080	26	2	0.08
	1090	29	2	0.09
	1100	3	1	0.01
	1110	27	2	0.09
	1120	33	13	0.11
	1130	23	2	0.07
	1140	20	4	0.07
	1150	215	18	0.70
	1160	37	3	0.12
	1170	52	6	0.17
	1180	32	2	0.10
	1190	85	3	0.28
	1200	108	25	0.35
	1210	13	2	0.04
	1220	49	16	0.16
	1230	14	2	0.05
	1240	50	8	0.16
	1250	76	10	0.25
	1260	99	14	0.32
	1270	89	41	0.29
	1280	110	30	0.36
	1290	23	3	0.07
	1300	53	9	0.17
	1310	44	4	0.14
	1320	8	3	0.03
	1330	123	21	0.40

	1340	9	2	0.03
	1350	47	14	0.15
	1360	88	25	0.29
	1370	107	20	0.35
	1380	39	3	0.13
	1390	42	4	0.14
	1400	16	1	0.05
	1410	36	4	0.12
	1420	37	7	0.12
	1430	6	2	0.02
	1440	69	35	0.22
	1450	15	2	0.05
	1460	33	7	0.11
	1470	6	2	0.02
	1480	15	5	0.05
	1490	124	42	0.40
	1500	14	2	0.05
	1510	8	3	0.03
	1520	41	2	0.13
	1530	22	3	0.07
	1540	9	1	0.03
	1550	13	1	0.04
	1560	14	1	0.05
	1570	17	5	0.06
	1580	17	3	0.06
	1590	5	1	0.02
	1600	27	3	0.09
	1610	12	2	0.04

	1620	14	5	0.05
	1630	6	2	0.02
	1640	19	3	0.06
	1650	11	2	0.04
	1660	34	7	0.11
	1670	25	3	0.08
	1680	20	4	0.07
	1690	18	4	0.06
	1700	7	2	0.02
	1710	10	1	0.03
	1720	21	4	0.07
	1730	4	1	0.01
	1740	46	20	0.15
	1750	11	2	0.04
	1760	17	2	0.06
	1770	3	1	0.01
	1780	17	1	0.06
	1790	9	4	0.03
	1800	7	4	0.02
	1810	6	1	0.02
	1820	6	2	0.02
	1830	6	2	0.02
	1840	8	1	0.03
	1850	5	1	0.02
	1860	2	1	0.01
	1870	2	1	0.01
	1880	4	2	0.01
	1890	3	1	0.01

	1900	9	5	0.03
	1910	1	1	0.00
	1920	5	3	0.02
	1930	1	1	0.00
	1940	1	1	0.00
	1950	1	1	0.00
	1960	4	1	0.01
	1970	1	1	0.00
	1980	1	1	0.00
	1990	2	2	0.01
	2000	9	4	0.03
	2010	8	2	0.03
	2020	9	3	0.03
	2030	4	3	0.01
	2040	2	1	0.01
	2050	3	1	0.01
	2060	3	1	0.01
	2070	1	1	0.00
	2080	3	1	0.01
	2090	4	1	0.01
	2110	1	1	0.00
	2120	2	2	0.01
	2130	13	2	0.04
	2140	6	2	0.02
	2150	1	1	0.00
	2160	3	1	0.01
	2170	5	1	0.02
	2180	2	2	0.01

	2190	1	1	0.00
	2200	1	1	0.00
	2210	1	1	0.00
	2220	3	3	0.01
	2230	1	1	0.00
	2240	1	1	0.00
	2250	1	1	0.00
	2260	1	1	0.00
	2270	1	1	0.00
	2280	1	1	0.00
	2290	3	1	0.01
	2300	4	2	0.01
	2310	17	5	0.06
	2320	13	7	0.04
	2330	6	3	0.02
	2340	1	1	0.00
	2350	12	2	0.04
	2360	1	1	0.00
	2370	1	1	0.00
	2380	4	1	0.01
	2390	1	1	0.00
	2400	3	3	0.01
	2410	3	1	0.01
	2420	1	1	0.00
	2430	1	1	0.00
	2440	1	1	0.00
	2450	2	2	0.01
	2460	2	2	0.01

	3020	4	2	0.01
	3030	1	1	0.00
	3040	1	1	0.00
	3050	1	1	0.00
	3060	6	2	0.02
	3070	1	1	0.00
	3080	2	1	0.01
	3090	4	1	0.01
	3100	3	2	0.01
	3110	4	2	0.01
	3120	3	1	0.01
	3130	6	2	0.02
	3140	3	1	0.01
	3160	5	1	0.02
	3170	1	1	0.00
	3180	1	1	0.00
	3190	1	1	0.00
	3200	3	2	0.01
	3210	38	4	0.12
	3230	3	3	0.01
	3250	1	1	0.00
	3260	6	1	0.02
	3270	1	1	0.00
	3280	1	1	0.00
	3290	10	2	0.03
	3300	1	1	0.00
	3320	1	1	0.00
	3330	4	1	0.01

	3340	1	1	0.00
	3350	3	2	0.01
	3360	1	1	0.00
	3370	2	1	0.01
	3380	8	3	0.03
	3390	3	2	0.01
	3400	1	1	0.00
	3410	2	1	0.01
	3420	2	1	0.01
	3430	7	1	0.02
	3440	1	1	0.00
	3450	5	4	0.02
	3460	4	1	0.01
	3470	3	1	0.01
	3480	3	2	0.01
	3490	1	1	0.00
	3500	2	1	0.01
	3510	1	1	0.00
	3520	5	1	0.02
	3530	1	1	0.00
	3540	1	1	0.00
	3550	3	2	0.01
	3560	1	1	0.00
	3570	2	1	0.01
	3580	2	1	0.01
	3590	1	1	0.00
	3600	1	1	0.00
	3610	2	1	0.01

	3620	1	1	0.00
	3630	1	1	0.00
	3640	1	1	0.00
	3650	1	1	0.00
	3660	5	1	0.02
	3670	2	1	0.01
	3680	4	2	0.01
	3690	1	1	0.00
	3700	1	1	0.00
	3710	3	1	0.01
	3720	7	1	0.02
	3730	7	4	0.02
	3740	2	1	0.01
	3750	1	1	0.00
	3760	1	1	0.00
	3780	2	1	0.01
	3790	1	1	0.00
	3800	1	1	0.00
	3810	4	1	0.01
	3820	1	1	0.00
	3830	1	1	0.00
	3860	1	1	0.00
	3870	1	1	0.00
	3880	1	1	0.00
	3890	1	1	0.00
	3900	2	1	0.01
	3910	3	1	0.01
	3920	1	1	0.00

	3930	3	1	0.01
	3940	1	1	0.00
	3950	1	1	0.00
	3960	1	1	0.00
	3970	1	1	0.00
	3980	1	1	0.00
	3990	1	1	0.00
	4000	1	1	0.00
	4010	2	1	0.01
	4020	2	1	0.01
	4030	3	2	0.01
	4040	1	1	0.00
	4050	1	1	0.00
	4060	1	1	0.00
	4070	1	1	0.00
	4080	5	1	0.02
	4090	1	1	0.00
	4100	2	1	0.01
	4110	1	1	0.00
	4120	4	2	0.01
	4140	2	1	0.01
	4150	1	1	0.00
	4160	1	1	0.00
	4170	1	1	0.00
	4180	1	1	0.00
	4190	1	1	0.00
	4200	1	1	0.00
BI	1	16093	205	44.06
	2	3307	38	9.05

O L I P	3	2978	51	8.15
	4	1449	33	3.97
	5	323	9	0.88
	6	138	8	0.38
	7	1736	37	4.75
	8	98	4	0.27
	9	14	2	0.04
	10	601	3	1.65
	11	2037	44	5.58
	12	155	13	0.42
	13	734	23	2.01
	14	357	7	0.98
	15	114	5	0.31
	16	109	5	0.30
	17	102	13	0.28
	18	17	1	0.05
	19	37	5	0.10
	20	7	1	0.02
	21	1610	65	4.41
	22	494	31	1.35
	23	994	26	2.72
	24	557	47	1.52
	25	129	2	0.35
	26	113	5	0.31
	27	203	7	0.56
	30	4	3	0.01
	31	27	1	0.07
	34	2	1	0.01
	98	8	2	0.02
	99	1347	20	3.69

Representação das classes no Nível 4			
	Class	Samples	Representation (%)

C A T H	1	3	0.01
	10	12700	41.32
	100	381	1.24
	1000	162	0.53
	10010	1	0.00
	10050	2	0.01
	10070	1	0.00
	10090	10	0.03
	1010	40	0.13
	10110	1	0.00
	10130	7	0.02
	10140	11	0.04
	10150	4	0.01
	10160	1	0.00
	10170	7	0.02
	10180	1	0.00
	10190	53	0.17
	1020	15	0.05
	10210	5	0.02
	10220	1	0.00
	10230	3	0.01
	10240	7	0.02
	10260	4	0.01
	10280	1	0.00
	102r8rA00	1	0.00
	1030	15	0.05
	10300	4	0.01
	10310	3	0.01

	10320	7	0.02
	10330	11	0.04
	10350	1	0.00
	10360	1	0.00
	10380	4	0.01
	10390	1	0.00
	103p4tA03	1	0.00
	1040	14	0.05
	10400	1	0.00
	10420	5	0.02
	10440	5	0.02
	10470	6	0.02
	10480	2	0.01
	10490	50	0.16
	105	2	0.01
	1050	11	0.04
	10540	3	0.01
	10550	1	0.00
	10580	5	0.02
	10590	2	0.01
	1060	31	0.10
	10600	1	0.00
	10610	5	0.02
	10620	1	0.00
	10630	1	0.00
	10640	2	0.01
	10660	2	0.01
	10670	1	0.00

	10680	1	0.00
	10690	1	0.00
	1070	14	0.05
	10700	2	0.01
	10710	3	0.01
	10720	1	0.00
	10730	1	0.00
	10740	5	0.02
	10750	1	0.00
	10760	2	0.01
	10770	2	0.01
	10780	1	0.00
	10790	4	0.01
	1080	25	0.08
	10800	4	0.01
	10810	4	0.01
	10820	1	0.00
	10830	1	0.00
	10840	1	0.00
	10850	1	0.00
	10860	22	0.07
	10870	1	0.00
	10880	2	0.01
	10890	4	0.01
	1090	19	0.06
	10900	7	0.02
	10910	1	0.00
	10920	1	0.00

	10930	1	0.00
	10940	1	0.00
	10950	2	0.01
	10960	1	0.00
	10970	1	0.00
	10980	1	0.00
	10990	2	0.01
	11	32	0.10
	110	214	0.70
	1100	59	0.19
	11000	1	0.00
	11010	1	0.00
	11020	1	0.00
	11030	2	0.01
	11040	1	0.00
	11050	2	0.01
	11060	1	0.00
	11070	1	0.00
	11080	1	0.00
	11090	1	0.00
	1110	55	0.18
	11100	1	0.00
	11110	2	0.01
	11120	3	0.01
	11130	1	0.00
	11140	1	0.00
	11150	1	0.00
	11160	1	0.00

	11170	2	0.01
	11180	1	0.00
	11190	1	0.00
	1120	22	0.07
	11200	1	0.00
	11210	1	0.00
	11220	1	0.00
	11230	8	0.03
	11240	1	0.00
	11250	1	0.00
	11260	2	0.01
	11270	2	0.01
	11280	1	0.00
	11290	1	0.00
	1130	19	0.06
	11300	1	0.00
	11310	1	0.00
	11320	1	0.00
	11330	1	0.00
	11340	3	0.01
	11350	4	0.01
	11370	1	0.00
	11380	3	0.01
	11390	1	0.00
	1140	15	0.05
	11400	1	0.00
	11410	1	0.00
	11420	1	0.00

	11440	2	0.01
	11450	3	0.01
	11460	2	0.01
	11480	1	0.00
	11490	1	0.00
	1150	16	0.05
	11500	1	0.00
	11510	1	0.00
	11530	1	0.00
	11540	1	0.00
	11550	2	0.01
	11570	1	0.00
	11580	3	0.01
	11590	1	0.00
	1160	14	0.05
	11600	1	0.00
	11610	1	0.00
	11620	1	0.00
	11630	1	0.00
	11650	1	0.00
	11660	1	0.00
	11670	1	0.00
	11680	1	0.00
	11690	1	0.00
	1170	25	0.08
	11700	1	0.00
	11710	3	0.01
	11720	1	0.00

	11730	1	0.00
	11740	1	0.00
	11750	1	0.00
	11760	1	0.00
	11770	1	0.00
	11780	1	0.00
	11790	1	0.00
	1180	135	0.44
	11800	1	0.00
	11810	1	0.00
	11820	1	0.00
	11830	1	0.00
	11840	1	0.00
	11850	1	0.00
	11860	1	0.00
	11880	1	0.00
	11890	1	0.00
	1190	16	0.05
	11900	1	0.00
	11920	1	0.00
	11930	1	0.00
	11940	1	0.00
	11950	1	0.00
	11960	1	0.00
	11970	1	0.00
	11980	2	0.01
	11990	1	0.00
	12	33	0.11

	120	253	0.82
	1200	10	0.03
	12000	1	0.00
	12020	1	0.00
	12030	1	0.00
	12050	1	0.00
	12060	1	0.00
	12080	1	0.00
	12090	2	0.01
	1210	14	0.05
	12100	2	0.01
	12110	1	0.00
	12120	1	0.00
	12140	1	0.00
	12150	1	0.00
	12160	1	0.00
	12170	2	0.01
	12180	1	0.00
	12190	1	0.00
	1220	49	0.16
	12210	1	0.00
	12220	1	0.00
	12230	1	0.00
	12240	2	0.01
	12250	2	0.01
	12260	2	0.01
	12270	1	0.00
	12280	1	0.00

	12290	1	0.00
	1230	29	0.09
	12300	1	0.00
	12310	1	0.00
	12320	1	0.00
	12330	1	0.00
	12350	1	0.00
	12360	1	0.00
	12370	4	0.01
	12380	1	0.00
	12390	1	0.00
	1240	51	0.17
	12420	1	0.00
	12430	1	0.00
	12440	1	0.00
	12450	1	0.00
	12470	1	0.00
	12480	1	0.00
	1250	8	0.03
	12500	4	0.01
	12520	1	0.00
	12530	1	0.00
	12540	1	0.00
	12550	1	0.00
	12570	1	0.00
	12580	1	0.00
	12590	1	0.00
	1260	16	0.05

	12600	1	0.00
	12610	3	0.01
	12620	1	0.00
	12630	1	0.00
	12640	1	0.00
	12650	2	0.01
	12660	1	0.00
	12670	1	0.00
	12690	1	0.00
	1270	8	0.03
	12700	1	0.00
	12710	1	0.00
	12740	2	0.01
	12760	1	0.00
	12780	33	0.11
	1280	17	0.06
	1290	20	0.07
	130	125	0.41
	1300	8	0.03
	1310	6	0.02
	1320	15	0.05
	1330	6	0.02
	1340	6	0.02
	1350	13	0.04
	1360	32	0.10
	1370	33	0.11
	1380	11	0.04
	1390	15	0.05

	140	471	1.53
	1400	28	0.09
	141	6	0.02
	1410	6	0.02
	1420	5	0.02
	1430	10	0.03
	1440	17	0.06
	1450	17	0.06
	1460	25	0.08
	1470	13	0.04
	1480	9	0.03
	1490	14	0.05
	150	494	1.61
	1500	5	0.02
	1510	13	0.04
	1520	11	0.04
	1530	8	0.03
	1540	8	0.03
	1550	4	0.01
	1560	10	0.03
	1570	6	0.02
	1580	29	0.09
	1590	7	0.02
	160	77	0.25
	1600	8	0.03
	1610	8	0.03
	1620	11	0.04
	1630	5	0.02

	1640	5	0.02
	1650	5	0.02
	1660	8	0.03
	1670	9	0.03
	1680	6	0.02
	1690	8	0.03
	170	188	0.61
	1700	19	0.06
	1710	10	0.03
	1720	5	0.02
	1730	22	0.07
	1740	8	0.03
	1750	4	0.01
	1760	18	0.06
	1770	6	0.02
	1780	8	0.03
	1790	5	0.02
	180	107	0.35
	1800	8	0.03
	1810	9	0.03
	1820	267	0.87
	1830	5	0.02
	1840	7	0.02
	1850	17	0.06
	1860	29	0.09
	1870	3	0.01
	1880	10	0.03
	1890	9	0.03

	190	84	0.27
	1900	10	0.03
	1910	15	0.05
	1920	7	0.02
	1930	16	0.05
	1940	13	0.04
	1950	15	0.05
	1960	4	0.01
	1970	24	0.08
	1980	107	0.35
	1990	3	0.01
	20	2032	6.61
	200	194	0.63
	2000	109	0.35
	2010	6	0.02
	2020	48	0.16
	20201wd5A02	1	0.00
	2030	15	0.05
	2040	5	0.02
	2050	6	0.02
	2060	13	0.04
	2070	5	0.02
	2080	6	0.02
	2090	5	0.02
	210	70	0.23
	2100	5	0.02
	2110	4	0.01
	2120	4	0.01

	2130	5	0.02
	2140	5	0.02
	2150	4	0.01
	2160	4	0.01
	2170	5	0.02
	2180	4	0.01
	2190	5	0.02
	220	71	0.23
	2200	4	0.01
	2210	5	0.02
	2220	8	0.03
	2230	5	0.02
	2240	4	0.01
	2250	4	0.01
	2260	3	0.01
	2270	3	0.01
	2280	3	0.01
	2290	3	0.01
	230	66	0.21
	2300	419	1.36
	2310	2	0.01
	2320	4	0.01
	2330	3	0.01
	2340	5	0.02
	2350	2	0.01
	2360	3	0.01
	2370	7	0.02
	2380	4	0.01

	2390	3	0.01
	24	17	0.06
	240	103	0.34
	2400	4	0.01
	2410	3	0.01
	2420	4	0.01
	2430	3	0.01
	2440	4	0.01
	2450	6	0.02
	2460	3	0.01
	2470	5	0.02
	2480	4	0.01
	2490	4	0.01
	250	60	0.20
	2500	4	0.01
	2510	4	0.01
	2520	3	0.01
	2530	4	0.01
	2540	3	0.01
	2550	3	0.01
	2560	3	0.01
	2570	7	0.02
	2580	8	0.03
	2590	3	0.01
	260	231	0.75
	2600	5	0.02
	261	9	0.03
	2610	3	0.01

	2620	8	0.03
	2630	6	0.02
	2640	1	0.00
	2650	3	0.01
	2660	6	0.02
	2670	3	0.01
	2680	1	0.00
	2690	4	0.01
	270	77	0.25
	2700	5	0.02
	2710	2	0.01
	2720	2	0.01
	2730	3	0.01
	2740	2	0.01
	2750	3	0.01
	2760	3	0.01
	2770	4	0.01
	2780	2	0.01
	280	59	0.19
	2800	3	0.01
	2810	3	0.01
	2820	3	0.01
	2830	5	0.02
	2840	4	0.01
	2850	3	0.01
	2860	4	0.01
	2870	2	0.01
	2880	2	0.01

	2890	3	0.01
	290	52	0.17
	2900	2	0.01
	2910	3	0.01
	2920	3	0.01
	2930	1	0.00
	2940	2	0.01
	2950	2	0.01
	2960	2	0.01
	2970	2	0.01
	2980	2	0.01
	2990	2	0.01
	30	1118	3.64
	300	599	1.95
	3000	4	0.01
	3010	1	0.00
	3020	2	0.01
	3030	2	0.01
	3040	2	0.01
	3050	8	0.03
	3060	1	0.00
	3070	2	0.01
	3080	4	0.01
	3090	2	0.01
	310	49	0.16
	3100	2	0.01
	3110	2	0.01
	3120	3	0.01

	3130	5	0.02
	3140	1	0.00
	3150	2	0.01
	3160	2	0.01
	3170	3	0.01
	3180	2	0.01
	3190	4	0.01
	32	4	0.01
	320	38	0.12
	3200	2	0.01
	3210	2	0.01
	3220	2	0.01
	3230	2	0.01
	3240	1	0.00
	3250	2	0.01
	3260	2	0.01
	3270	2	0.01
	3280	1	0.00
	3290	7	0.02
	330	226	0.74
	3300	1	0.00
	3310	2	0.01
	3320	2	0.01
	3330	2	0.01
	3340	2	0.01
	3350	2	0.01
	3360	2	0.01
	3370	2	0.01

	3380	2	0.01
	3390	1	0.00
	340	61	0.20
	3400	2	0.01
	3410	2	0.01
	3420	2	0.01
	3430	4	0.01
	3440	1	0.00
	3450	2	0.01
	3460	2	0.01
	3470	2	0.01
	3480	2	0.01
	3490	1	0.00
	350	43	0.14
	3500	1	0.00
	3510	2	0.01
	3530	2	0.01
	3540	2	0.01
	3550	1	0.00
	3570	1	0.00
	3580	1	0.00
	3590	1	0.00
	360	116	0.38
	3600	1	0.00
	3610	1	0.00
	3620	3	0.01
	3630	1	0.00
	3640	1	0.00

	3650	1	0.00
	3670	1	0.00
	3680	1	0.00
	3690	4	0.01
	370	62	0.20
	3700	1	0.00
	3710	2	0.01
	3720	1	0.00
	3730	1	0.00
	3740	1	0.00
	3750	1	0.00
	3760	1	0.00
	3770	1	0.00
	3780	1	0.00
	3790	1	0.00
	380	50	0.16
	3800	1	0.00
	3810	1	0.00
	3820	1	0.00
	3830	1	0.00
	3850	1	0.00
	3860	1	0.00
	3870	1	0.00
	3880	2	0.01
	3890	1	0.00
	390	66	0.21
	3900	1	0.00
	3910	3	0.01

	3920	1	0.00
	3930	1	0.00
	3940	1	0.00
	3950	1	0.00
	3960	2	0.01
	3970	1	0.00
	3980	1	0.00
	40	823	2.68
	400	32	0.10
	4000	1	0.00
	4020	1	0.00
	4030	2	0.01
	4040	1	0.00
	4050	1	0.00
	4060	1	0.00
	4070	2	0.01
	41	2	0.01
	410	56	0.18
	4100	3	0.01
	4110	1	0.00
	4120	1	0.00
	4130	1	0.00
	4140	1	0.00
	4150	3	0.01
	4160	1	0.00
	4170	1	0.00
	4180	1	0.00
	420	139	0.45

	430	38	0.12
	44	5	0.02
	440	42	0.14
	450	76	0.25
	460	48	0.16
	470	31	0.10
	480	26	0.08
	490	35	0.11
	50	564	1.83
	500	26	0.08
	510	38	0.12
	520	23	0.07
	530	28	0.09
	540	32	0.10
	550	24	0.08
	560	37	0.12
	570	25	0.08
	580	30	0.10
	590	25	0.08
	60	777	2.53
	600	27	0.09
	610	20	0.07
	620	189	0.61
	630	22	0.07
	640	27	0.09
	650	44	0.14
	660	32	0.10
	670	23	0.07

	680	30	0.10
	690	25	0.08
	70	487	1.58
	700	23	0.07
	710	24	0.08
	720	714	2.32
	730	27	0.09
	740	53	0.17
	750	16	0.05
	760	17	0.06
	770	24	0.08
	7700	3	0.01
	780	19	0.06
	790	57	0.19
	80	489	1.59
	800	35	0.11
	810	20	0.07
	820	18	0.06
	830	22	0.07
	840	17	0.06
	850	38	0.12
	860	15	0.05
	870	21	0.07
	880	77	0.25
	890	41	0.13
	90	344	1.12
	900	12	0.04
	910	9	0.03

	9100	2	0.01
	920	59	0.19
	9200	1	0.00
	930	19	0.06
	940	15	0.05
	950	12	0.04
	960	19	0.06
	970	72	0.23
	980	26	0.08
	990	18	0.06
BIOLIP	1	5237	14.34
	2	2064	5.65
	3	1809	4.95
	4	1654	4.53
	5	1229	3.36
	6	1533	4.20
	7	1177	3.22
	8	796	2.18
	9	818	2.24
	10	381	1.04
	11	620	1.70
	12	472	1.29
	13	423	1.16
	14	402	1.10
	15	408	1.12
	16	806	2.21
	17	770	2.11
	18	785	2.15

	19	288	0.79
	20	274	0.75
	21	516	1.41
	22	465	1.27
	23	374	1.02
	24	437	1.20
	25	229	0.63
	26	131	0.36
	27	355	0.97
	28	203	0.56
	29	86	0.24
	30	293	0.80
	31	119	0.33
	32	76	0.21
	33	169	0.46
	34	147	0.40
	35	147	0.40
	36	164	0.45
	37	534	1.46
	38	116	0.32
	39	550	1.51
	40	192	0.53
	41	115	0.31
	42	143	0.39
	43	133	0.36
	44	48	0.13
	45	203	0.56
	46	334	0.91

	47	87	0.24
	48	477	1.31
	49	187	0.51
	50	58	0.16
	51	33	0.09
	52	151	0.41
	53	25	0.07
	54	52	0.14
	55	75	0.21
	56	63	0.17
	57	60	0.16
	58	48	0.13
	59	83	0.23
	60	88	0.24
	61	27	0.07
	62	103	0.28
	63	36	0.10
	64	58	0.16
	65	58	0.16
	66	18	0.05
	67	10	0.03
	68	29	0.08
	69	56	0.15
	70	70	0.19
	71	36	0.10
	72	59	0.16
	73	86	0.24
	74	110	0.30

	75	17	0.05
	76	3	0.01
	77	19	0.05
	78	75	0.21
	79	20	0.05
	80	15	0.04
	81	23	0.06
	82	26	0.07
	83	24	0.07
	84	37	0.10
	85	37	0.10
	86	35	0.10
	87	10	0.03
	88	86	0.24
	89	21	0.06
	90	26	0.07
	91	97	0.27
	92	79	0.22
	93	25	0.07
	94	17	0.05
	95	78	0.21
	96	20	0.05
	97	73	0.20
	98	66	0.18
	99	32	0.09
	100	51	0.14
	101	25	0.07
	102	8	0.02

	103	21	0.06
	104	14	0.04
	105	7	0.02
	107	11	0.03
	108	6	0.02
	109	6	0.02
	110	4	0.01
	111	4	0.01
	112	155	0.42
	113	31	0.08
	114	19	0.05
	115	3	0.01
	117	22	0.06
	118	3	0.01
	119	1	0.00
	122	1	0.00
	123	10	0.03
	125	18	0.05
	126	1	0.00
	127	23	0.06
	128	6	0.02
	129	22	0.06
	130	11	0.03
	131	3	0.01
	132	6	0.02
	133	13	0.04
	135	29	0.08
	136	6	0.02

	137	11	0.03
	138	4	0.01
	139	11	0.03
	141	4	0.01
	143	10	0.03
	144	9	0.02
	145	19	0.05
	146	87	0.24
	147	7	0.02
	148	55	0.15
	149	3	0.01
	150	2	0.01
	151	20	0.05
	152	3	0.01
	153	61	0.17
	154	4	0.01
	156	2	0.01
	157	5	0.01
	158	23	0.06
	159	6	0.02
	161	4	0.01
	163	1	0.00
	164	2	0.01
	168	14	0.04
	169	13	0.04
	173	5	0.01
	174	1	0.00
	175	19	0.05

	177	1	0.00
	178	7	0.02
	179	4	0.01
	180	12	0.03
	182	5	0.01
	184	9	0.02
	186	19	0.05
	188	4	0.01
	192	4	0.01
	193	5	0.01
	195	6	0.02
	199	1	0.00
	202	2	0.01
	203	3	0.01
	204	36	0.10
	205	42	0.11
	207	2	0.01
	209	2	0.01
	211	7	0.02
	213	4	0.01
	217	14	0.04
	218	3	0.01
	219	11	0.03
	221	5	0.01
	222	1	0.00
	227	1	0.00
	228	4	0.01
	236	10	0.03

	237	1	0.00
	242	1	0.00
	247	4	0.01
	248	1	0.00
	251	1	0.00
	252	4	0.01
	253	9	0.02
	255	8	0.02
	262	3	0.01
	263	4	0.01
	267	41	0.11
	268	3	0.01
	271	4	0.01
	272	6	0.02
	274	1	0.00
	275	6	0.02
	276	2	0.01
	289	6	0.02
	290	3	0.01
	-	3820	10.46

ANEXO C - REPRESENTAÇÃO DAS CARACTERÍSTICAS DO *CATH* E *BIOLIP*

	CATH	**Biolip**
Features	21	A
	19	C
	17	D
	15	E
	13	F
	11	G
	7	H
	5	I
	3	K
	1	L
		M
		N
		P
		Q
		R
		S
		T
		V
		W
		Y
		AA
		AC
		AD
		AE
		AF
		AG
		AH
		AI

		AK
		AL
		AM
		AN
		AP
		AQ
		AR
		AS
		AT
		AV
		AW
		AY
		CA
		CC
		CD
		CE
		CF
		CG
		CH
		CI
		CK
		CL
		CM
		CN
		CP
		CQ
		CR
		CS
		CT
		CV
		CW
		CY
		DA

		DC
		DD
		DE
		DF
		DG
		DH
		DI
		DK
		DL
		DM
		DN
		DP
		DQ
		DR
		DS
		DT
		DV
		DW
		DY
		EA
		EC
		ED
		EE
		EF
		EG
		EH
		EI
		EK
		EL
		EM
		EN
		EP
		EQ

		ER
		ES
		ET
		EV
		EW
		EY
		FA
		FC
		FD
		FE
		FF
		FG
		FH
		FI
		FK
		FL
		FM
		FN
		FP
		FQ
		FR
		FS
		FT
		FV
		FW
		FY
		GA
		GC
		GD
		GE
		GF
		GG
		GH

		GI
		GK
		GL
		GM
		GN
		GP
		GQ
		GR
		GS
		GT
		GV
		GW
		GY
		HA
		HC
		HD
		HE
		HF
		HG
		HH
		HI
		HK
		HL
		HM
		HN
		HP
		HQ
		HR
		HS
		HT
		HV
		HW
		HY

		IA
		IC
		ID
		IE
		IF
		IG
		IH
		II
		IK
		IL
		IM
		IN
		IP
		IQ
		IR
		IS
		IT
		IV
		IW
		IY
		KA
		KC
		KD
		KE
		KF
		KG
		KH
		KI
		KK
		KL
		KM
		KN
		KP

		KQ
		KR
		KS
		KT
		KV
		KW
		KY
		LA
		LC
		LD
		LE
		LF
		LG
		LH
		LI
		LK
		LL
		LM
		LN
		LP
		LQ
		LR
		LS
		LT
		LV
		LW
		LY
		MA
		MC
		MD
		ME
		MF
		MG

		MH
		MI
		MK
		ML
		MM
		MN
		MP
		MQ
		MR
		MS
		MT
		MV
		MW
		MY
		NA
		NC
		ND
		NE
		NF
		NG
		NH
		NI
		NK
		NL
		NM
		NN
		NP
		NQ
		NR
		NS
		NT
		NV
		NW

		NY
		PA
		PC
		PD
		PE
		PF
		PG
		PH
		PI
		PK
		PL
		PM
		PN
		PP
		PQ
		PR
		PS
		PT
		PV
		PW
		PY
		QA
		QC
		QD
		QE
		QF
		QG
		QH
		QI
		QK
		QL
		QM
		QN

		QP
		QQ
		QR
		QS
		QT
		QV
		QW
		QY
		RA
		RC
		RD
		RE
		RF
		RG
		RH
		RI
		RK
		RL
		RM
		RN
		RP
		RQ
		RR
		RS
		RT
		RV
		RW
		RY
		SA
		SC
		SD
		SE
		SF

		SG
		SH
		SI
		SK
		SL
		SM
		SN
		SP
		SQ
		SR
		SS
		ST
		SV
		SW
		SY
		TA
		TC
		TD
		TE
		TF
		TG
		TH
		TI
		TK
		TL
		TM
		TN
		TP
		TQ
		TR
		TS
		TT
		TV

		TW
		TY
		VA
		VC
		VD
		VE
		VF
		VG
		VH
		VI
		VK
		VL
		VM
		VN
		VP
		VQ
		VR
		VS
		VT
		VV
		VW
		VY
		WA
		WC
		WD
		WE
		WF
		WG
		WH
		WI
		WK
		WL
		WM

		WN
		WP
		WQ
		WR
		WS
		WT
		WV
		WW
		WY
		YA
		YC
		YD
		YE
		YF
		YG
		YH
		YI
		YK
		YL
		YM
		YN
		YP
		YQ
		YR
		YS
		YT
		YV
		YW
		YY

ANEXO D - REPRESENTAÇÃO DAS CLASSES POR NÍVEL DEPOIS DO BALANCEAMENTO

		1º Level		2º Level		3º Level		4º Level	
		CATH	BIOLIP	CATH	BIOLIP	CATH	BIOLIP	CATH	BIOLIP
Classes	per level	4	6	23	22	324	29	371	206
	per node	-	-	17	15	128	17	52	46
Samples	mean	3,845	4,374	669	1,191	47	902	41	127
	std	3,298	3,740	1,289	1,859	262	2,613	122	225
	vc	0.86	0.86	1.93	1.56	5.52	2.90	2.94	1.76

Representação das classes no Nível 1				
	Class	Samples	Number of labels	Representation (%)
CATH	B	3757	5	24.43
	C	3516	21	22.86
	D	8080	14	52.54
	E	26	1	0.17
BIOLIP	A	5226	22	19.91
	B	8454	9	32.22
	C	8952	11	34.11
	D	2187	6	8.33
	E	753	6	2.87
	F	670	6	2.55

Representação das classes no Nível 2				
	Class	Samples	Number of labels	Representation (%)
CATH	B	3757	5	24.43
	C	3516	21	22.86
	D	8080	14	52.54
	E	26	1	0.17
B	A	5226	22	19.91

IOLIP	B	8454	9	32.22
	C	8952	11	34.11
	D	2187	6	8.33
	E	753	6	2.87
	F	670	6	2.55
BIOLIP	A	6017	17	22.93
	B	3928	7	14.97
	C	1764	8	6.72
	D	5026	20	19.15
	E	2236	7	8.52
	F	1330	9	5.07
	G	4017	14	15.31
	H	245	5	0.93
	I	73	1	0.28
	J	48	3	0.18
	K	207	1	0.79
	L	7	2	0.03
	M	242	3	0.92
	N	866	13	3.30
	O	45	1	0.17
	P	9	2	0.03
	Q	50	2	0.19
	R	62	4	0.24
	S	1	1	0.00
	T	7	1	0.03
	U	5	1	0.02
	V	26	1	0.10

Representação das classes no Nível 3				
	Class	Samples	Number of labels	Representation (%)
CATH	4	2	2	0.01
	5	423	99	2.75
	8	201	114	1.31

	9	15	7	0.10
	10	780	247	5.07
	12	9	7	0.06
	15	16	1	0.10
	16	1	1	0.01
	20	1386	70	9.01
	21	8	5	0.05
	22	1	1	0.01
	25	108	58	0.70
	27	1	1	0.01
	28	42	22	0.27
	29	57	22	0.37
	30	490	108	3.19
	31	2	2	0.01
	35	1	1	0.01
	36	4	3	0.03
	37	1	1	0.01
	38	1	1	0.01
	39	5	1	0.03
	40	1391	264	9.04
	42	14	1	0.09
	43	3	2	0.02
	45	1	1	0.01
	47	8	1	0.05
	50	3888	196	25.28
	54	1	1	0.01
	56	8	6	0.05
	58	400	189	2.60
	59	3	2	0.02
	60	51	23	0.33
	66	1	1	0.01
	69	1	1	0.01
	70	1052	209	6.84
	76	1	1	0.01
	79	17	1	0.11

	80	12	4	0.08
	81	1	1	0.01
	85	1	1	0.01
	87	1	1	0.01
	89	2	1	0.01
	90	7	2	0.05
	91	13	7	0.08
	98	22	10	0.14
	100	12	2	0.08
	101	1	1	0.01
	105	2	1	0.01
	109	12	1	0.08
	110	55	14	0.36
	120	1075	181	6.99
	128	108	58	0.70
	129	39	8	0.25
	130	19	4	0.12
	132	12	7	0.08
	140	39	14	0.25
	150	242	78	1.57
	155	4	1	0.03
	160	183	74	1.19
	170	7	3	0.05
	175	1	1	0.01
	180	31	7	0.20
	190	151	21	0.98
	199	1	1	0.01
	200	61	19	0.40
	209	1	1	0.01
	210	7	3	0.05
	215	1	1	0.01
	220	22	10	0.14
	225	4	2	0.03
	226	11	3	0.07
	228	3	1	0.02

	230	44	7	0.29
	238	45	21	0.29
	240	10	6	0.07
	245	3	1	0.02
	246	25	17	0.16
	250	9	1	0.06
	260	47	17	0.31
	270	13	3	0.08
	272	4	3	0.03
	274	12	6	0.08
	275	6	2	0.04
	280	5	1	0.03
	286	8	6	0.05
	287	358	94	2.33
	288	1	1	0.01
	290	10	5	0.07
	300	65	29	0.42
	309	2	1	0.01
	310	54	26	0.35
	320	2	2	0.01
	330	8	7	0.05
	340	7	4	0.05
	350	4	1	0.03
	357	46	12	0.30
	360	26	9	0.17
	365	1	1	0.01
	366	2	1	0.01
	367	1	1	0.01
	370	1	1	0.01
	375	2	1	0.01
	379	1	1	0.01
	380	2	2	0.01
	390	36	11	0.23
	400	3	1	0.02
	405	2	1	0.01

	410	4	1	0.03
	413	2	1	0.01
	418	6	4	0.04
	420	133	45	0.86
	428	3	2	0.02
	429	8	1	0.05
	430	6	4	0.04
	437	6	3	0.04
	439	1	1	0.01
	440	1	1	0.01
	443	1	1	0.01
	450	234	63	1.52
	460	17	8	0.11
	462	1	1	0.01
	465	4	2	0.03
	470	29	9	0.19
	472	26	13	0.17
	479	4	3	0.03
	480	1	1	0.01
	489	1	1	0.01
	490	16	12	0.10
	497	1	1	0.01
	499	1	1	0.01
	500	4	1	0.03
	505	4	1	0.03
	506	2	1	0.01
	510	27	1	0.18
	520	2	1	0.01
	530	25	7	0.16
	532	1	1	0.01
	533	11	3	0.07
	540	9	2	0.06
	550	19	2	0.12
	555	3	1	0.02
	559	6	3	0.04

	560	1	1	0.01
	565	10	2	0.07
	570	2	1	0.01
	572	1	1	0.01
	579	1	1	0.01
	580	10	1	0.07
	590	1	1	0.01
	600	6	1	0.04
	605	4	1	0.03
	620	3	2	0.02
	630	73	14	0.47
	640	38	2	0.25
	650	7	1	0.05
	660	4	2	0.03
	700	2	2	0.01
	710	9	1	0.06
	718	2	1	0.01
	720	38	10	0.25
	730	6	1	0.04
	740	1	1	0.01
	750	46	16	0.30
	760	10	2	0.07
	780	5	2	0.03
	800	2	1	0.01
	810	2	1	0.01
	830	12	1	0.08
	840	1	1	0.01
	850	4	1	0.03
	860	2	2	0.01
	870	2	2	0.01
	890	8	7	0.05
	900	3	2	0.02
	910	5	1	0.03
	920	10	6	0.07
	930	20	6	0.13

	940	3	2	0.02
	950	4	1	0.03
	960	4	4	0.03
	970	3	3	0.02
	980	3	3	0.02
	990	2	1	0.01
	1000	6	3	0.04
	1010	9	2	0.06
	1040	8	5	0.05
	1050	30	11	0.20
	1060	1	1	0.01
	1070	13	3	0.08
	1080	3	2	0.02
	1090	4	1	0.03
	1110	5	1	0.03
	1120	17	10	0.11
	1130	2	1	0.01
	1140	6	1	0.04
	1150	61	14	0.40
	1160	7	2	0.05
	1170	5	3	0.03
	1180	5	1	0.03
	1190	7	2	0.05
	1200	44	21	0.29
	1210	2	1	0.01
	1220	22	13	0.14
	1230	2	2	0.01
	1240	11	5	0.07
	1250	12	8	0.08
	1260	27	10	0.18
	1270	58	37	0.38
	1280	45	27	0.29
	1290	4	1	0.03
	1300	13	4	0.08
	1310	9	4	0.06

	1330	49	20	0.32
	1340	1	1	0.01
	1350	18	10	0.12
	1360	50	22	0.33
	1370	36	17	0.23
	1380	7	3	0.05
	1390	6	3	0.04
	1400	3	1	0.02
	1410	10	1	0.07
	1420	9	6	0.06
	1430	2	2	0.01
	1440	52	32	0.34
	1450	3	2	0.02
	1460	3	3	0.02
	1470	1	1	0.01
	1480	1	1	0.01
	1490	84	39	0.55
	1500	2	2	0.01
	1510	3	2	0.02
	1520	6	1	0.04
	1530	2	2	0.01
	1540	2	1	0.01
	1550	1	1	0.01
	1560	1	1	0.01
	1570	2	1	0.01
	1580	6	2	0.04
	1590	1	1	0.01
	1600	2	1	0.01
	1610	1	1	0.01
	1620	3	2	0.02
	1630	1	1	0.01
	1640	4	2	0.03
	1650	4	1	0.03
	1660	6	3	0.04
	1670	6	3	0.04

	1680	4	1	0.03
	1690	7	2	0.05
	1700	2	1	0.01
	1710	3	1	0.02
	1720	5	3	0.03
	1730	3	1	0.02
	1740	23	19	0.15
	1750	2	1	0.01
	1780	1	1	0.01
	1820	1	1	0.01
	1840	3	1	0.02
	1850	2	1	0.01
	1860	1	1	0.01
	1870	1	1	0.01
	1890	2	1	0.01
	1900	2	2	0.01
	1920	3	2	0.02
	1930	1	1	0.01
	1940	1	1	0.01
	1970	1	1	0.01
	2000	2	2	0.01
	2010	3	2	0.02
	2020	2	2	0.01
	2030	1	1	0.01
	2060	1	1	0.01
	2080	2	1	0.01
	2120	1	1	0.01
	2130	1	1	0.01
	2170	1	1	0.01
	2200	1	1	0.01
	2210	1	1	0.01
	2220	1	1	0.01
	2310	2	2	0.01
	2320	4	3	0.03
	2380	2	1	0.01

	2400	1	1	0.01
	2410	1	1	0.01
	2440	1	1	0.01
	3090	2	1	0.01
	3100	1	1	0.01
	3120	1	1	0.01
	3130	1	1	0.01
	3140	1	1	0.01
	3200	1	1	0.01
	3210	2	1	0.01
	3260	1	1	0.01
	3290	1	1	0.01
	3330	1	1	0.01
	3340	1	1	0.01
	3380	3	3	0.02
	3410	1	1	0.01
	3420	1	1	0.01
	3430	1	1	0.01
	3490	1	1	0.01
	3500	1	1	0.01
	3510	1	1	0.01
	3540	1	1	0.01
	3550	2	1	0.01
	3580	2	1	0.01
	3590	1	1	0.01
	3600	1	1	0.01
	3620	1	1	0.01
	3630	1	1	0.01
	3640	1	1	0.01
	3660	2	1	0.01
	3710	1	1	0.01
	3720	1	1	0.01
	3730	1	1	0.01
	3860	1	1	0.01
	3920	1	1	0.01

	3940	1	1	0.01
	3950	1	1	0.01
	4080	2	1	0.01
	4110	1	1	0.01
	4120	1	1	0.01
	4190	1	1	0.01
BIOLIP	1	14140	205	53.88
	2	2039	38	7.77
	3	1961	51	7.47
	4	1092	32	4.16
	5	67	7	0.26
	6	69	7	0.26
	7	952	37	3.63
	8	20	4	0.08
	10	6	2	0.02
	11	1147	43	4.37
	12	95	13	0.36
	13	645	22	2.46
	14	59	4	0.22
	15	32	2	0.12
	16	22	3	0.08
	17	65	12	0.25
	18	1	1	0.00
	19	20	4	0.08
	20	5	1	0.02
	21	1005	64	3.83
	22	365	30	1.39
	23	1113	23	4.24
	24	537	45	2.05
	25	13	2	0.05
	26	33	4	0.13
	27	57	6	0.22
	30	1	1	0.00
	31	3	1	0.01
	99	581	20	2.21

Representação das classes no Nível 4			
	Class	Samples	Representation (%)
C A T H	1	3	0.01
	10	12700	41.32
	100	381	1.24
	1000	162	0.53
	10010	1	0.00
	10050	2	0.01
	10070	1	0.00
	10090	10	0.03
	1010	40	0.13
	10110	1	0.00
	10130	7	0.02
	10140	11	0.04
	10150	4	0.01
	10160	1	0.00
	10170	7	0.02
	10180	1	0.00
	10190	53	0.17
	1020	15	0.05
	10210	5	0.02
	10220	1	0.00
	10230	3	0.01
	10240	7	0.02
	10260	4	0.01
	10280	1	0.00
	102r8rA00	1	0.00
	1030	15	0.05
	10300	4	0.01
	10310	3	0.01
	10320	7	0.02
	10330	11	0.04
	10350	1	0.00

	10360	1	0.00
	10380	4	0.01
	10390	1	0.00
	103p4tA03	1	0.00
	1040	14	0.05
	10400	1	0.00
	10420	5	0.02
	10440	5	0.02
	10470	6	0.02
	10480	2	0.01
	10490	50	0.16
	105	2	0.01
	1050	11	0.04
	10540	3	0.01
	10550	1	0.00
	10580	5	0.02
	10590	2	0.01
	1060	31	0.10
	10600	1	0.00
	10610	5	0.02
	10620	1	0.00
	10630	1	0.00
	10640	2	0.01
	10660	2	0.01
	10670	1	0.00
	10680	1	0.00
	10690	1	0.00
	1070	14	0.05
	10700	2	0.01
	10710	3	0.01
	10720	1	0.00
	10730	1	0.00
	10740	5	0.02
	10750	1	0.00
	10760	2	0.01

	10770	2	0.01
	10780	1	0.00
	10790	4	0.01
	1080	25	0.08
	10800	4	0.01
	10810	4	0.01
	10820	1	0.00
	10830	1	0.00
	10840	1	0.00
	10850	1	0.00
	10860	22	0.07
	10870	1	0.00
	10880	2	0.01
	10890	4	0.01
	1090	19	0.06
	10900	7	0.02
	10910	1	0.00
	10920	1	0.00
	10930	1	0.00
	10940	1	0.00
	10950	2	0.01
	10960	1	0.00
	10970	1	0.00
	10980	1	0.00
	10990	2	0.01
	11	32	0.10
	110	214	0.70
	1100	59	0.19
	11000	1	0.00
	11010	1	0.00
	11020	1	0.00
	11030	2	0.01
	11040	1	0.00
	11050	2	0.01
	11060	1	0.00

	11070	1	0.00
	11080	1	0.00
	11090	1	0.00
	1110	55	0.18
	11100	1	0.00
	11110	2	0.01
	11120	3	0.01
	11130	1	0.00
	11140	1	0.00
	11150	1	0.00
	11160	1	0.00
	11170	2	0.01
	11180	1	0.00
	11190	1	0.00
	1120	22	0.07
	11200	1	0.00
	11210	1	0.00
	11220	1	0.00
	11230	8	0.03
	11240	1	0.00
	11250	1	0.00
	11260	2	0.01
	11270	2	0.01
	11280	1	0.00
	11290	1	0.00
	1130	19	0.06
	11300	1	0.00
	11310	1	0.00
	11320	1	0.00
	11330	1	0.00
	11340	3	0.01
	11350	4	0.01
	11370	1	0.00
	11380	3	0.01
	11390	1	0.00

	1140	15	0.05
	11400	1	0.00
	11410	1	0.00
	11420	1	0.00
	11440	2	0.01
	11450	3	0.01
	11460	2	0.01
	11480	1	0.00
	11490	1	0.00
	1150	16	0.05
	11500	1	0.00
	11510	1	0.00
	11530	1	0.00
	11540	1	0.00
	11550	2	0.01
	11570	1	0.00
	11580	3	0.01
	11590	1	0.00
	1160	14	0.05
	11600	1	0.00
	11610	1	0.00
	11620	1	0.00
	11630	1	0.00
	11650	1	0.00
	11660	1	0.00
	11670	1	0.00
	11680	1	0.00
	11690	1	0.00
	1170	25	0.08
	11700	1	0.00
	11710	3	0.01
	11720	1	0.00
	11730	1	0.00
	11740	1	0.00
	11750	1	0.00

	11760	1	0.00
	11770	1	0.00
	11780	1	0.00
	11790	1	0.00
	1180	135	0.44
	11800	1	0.00
	11810	1	0.00
	11820	1	0.00
	11830	1	0.00
	11840	1	0.00
	11850	1	0.00
	11860	1	0.00
	11880	1	0.00
	11890	1	0.00
	1190	16	0.05
	11900	1	0.00
	11920	1	0.00
	11930	1	0.00
	11940	1	0.00
	11950	1	0.00
	11960	1	0.00
	11970	1	0.00
	11980	2	0.01
	11990	1	0.00
	12	33	0.11
	120	253	0.82
	1200	10	0.03
	12000	1	0.00
	12020	1	0.00
	12030	1	0.00
	12050	1	0.00
	12060	1	0.00
	12080	1	0.00
	12090	2	0.01
	1210	14	0.05

	12100	2	0.01
	12110	1	0.00
	12120	1	0.00
	12140	1	0.00
	12150	1	0.00
	12160	1	0.00
	12170	2	0.01
	12180	1	0.00
	12190	1	0.00
	1220	49	0.16
	12210	1	0.00
	12220	1	0.00
	12230	1	0.00
	12240	2	0.01
	12250	2	0.01
	12260	2	0.01
	12270	1	0.00
	12280	1	0.00
	12290	1	0.00
	1230	29	0.09
	12300	1	0.00
	12310	1	0.00
	12320	1	0.00
	12330	1	0.00
	12350	1	0.00
	12360	1	0.00
	12370	4	0.01
	12380	1	0.00
	12390	1	0.00
	1240	51	0.17
	12420	1	0.00
	12430	1	0.00
	12440	1	0.00
	12450	1	0.00
	12470	1	0.00

	12480	1	0.00
	1250	8	0.03
	12500	4	0.01
	12520	1	0.00
	12530	1	0.00
	12540	1	0.00
	12550	1	0.00
	12570	1	0.00
	12580	1	0.00
	12590	1	0.00
	1260	16	0.05
	12600	1	0.00
	12610	3	0.01
	12620	1	0.00
	12630	1	0.00
	12640	1	0.00
	12650	2	0.01
	12660	1	0.00
	12670	1	0.00
	12690	1	0.00
	1270	8	0.03
	12700	1	0.00
	12710	1	0.00
	12740	2	0.01
	12760	1	0.00
	12780	33	0.11
	1280	17	0.06
	1290	20	0.07
	130	125	0.41
	1300	8	0.03
	1310	6	0.02
	1320	15	0.05
	1330	6	0.02
	1340	6	0.02
	1350	13	0.04

	1360	32	0.10
	1370	33	0.11
	1380	11	0.04
	1390	15	0.05
	140	471	1.53
	1400	28	0.09
	141	6	0.02
	1410	6	0.02
	1420	5	0.02
	1430	10	0.03
	1440	17	0.06
	1450	17	0.06
	1460	25	0.08
	1470	13	0.04
	1480	9	0.03
	1490	14	0.05
	150	494	1.61
	1500	5	0.02
	1510	13	0.04
	1520	11	0.04
	1530	8	0.03
	1540	8	0.03
	1550	4	0.01
	1560	10	0.03
	1570	6	0.02
	1580	29	0.09
	1590	7	0.02
	160	77	0.25
	1600	8	0.03
	1610	8	0.03
	1620	11	0.04
	1630	5	0.02
	1640	5	0.02
	1650	5	0.02
	1660	8	0.03

	1670	9	0.03
	1680	6	0.02
	1690	8	0.03
	170	188	0.61
	1700	19	0.06
	1710	10	0.03
	1720	5	0.02
	1730	22	0.07
	1740	8	0.03
	1750	4	0.01
	1760	18	0.06
	1770	6	0.02
	1780	8	0.03
	1790	5	0.02
	180	107	0.35
	1800	8	0.03
	1810	9	0.03
	1820	267	0.87
	1830	5	0.02
	1840	7	0.02
	1850	17	0.06
	1860	29	0.09
	1870	3	0.01
	1880	10	0.03
	1890	9	0.03
	190	84	0.27
	1900	10	0.03
	1910	15	0.05
	1920	7	0.02
	1930	16	0.05
	1940	13	0.04
	1950	15	0.05
	1960	4	0.01
	1970	24	0.08
	1980	107	0.35

	1990	3	0.01
	20	2032	6.61
	200	194	0.63
	2000	109	0.35
	2010	6	0.02
	2020	48	0.16
	20201wd5A02	1	0.00
	2030	15	0.05
	2040	5	0.02
	2050	6	0.02
	2060	13	0.04
	2070	5	0.02
	2080	6	0.02
	2090	5	0.02
	210	70	0.23
	2100	5	0.02
	2110	4	0.01
	2120	4	0.01
	2130	5	0.02
	2140	5	0.02
	2150	4	0.01
	2160	4	0.01
	2170	5	0.02
	2180	4	0.01
	2190	5	0.02
	220	71	0.23
	2200	4	0.01
	2210	5	0.02
	2220	8	0.03
	2230	5	0.02
	2240	4	0.01
	2250	4	0.01
	2260	3	0.01
	2270	3	0.01

	2280	3	0.01
	2290	3	0.01
	230	66	0.21
	2300	419	1.36
	2310	2	0.01
	2320	4	0.01
	2330	3	0.01
	2340	5	0.02
	2350	2	0.01
	2360	3	0.01
	2370	7	0.02
	2380	4	0.01
	2390	3	0.01
	24	17	0.06
	240	103	0.34
	2400	4	0.01
	2410	3	0.01
	2420	4	0.01
	2430	3	0.01
	2440	4	0.01
	2450	6	0.02
	2460	3	0.01
	2470	5	0.02
	2480	4	0.01
	2490	4	0.01
	250	60	0.20
	2500	4	0.01
	2510	4	0.01
	2520	3	0.01
	2530	4	0.01
	2540	3	0.01
	2550	3	0.01
	2560	3	0.01
	2570	7	0.02
	2580	8	0.03

	2590	3	0.01
	260	231	0.75
	2600	5	0.02
	261	9	0.03
	2610	3	0.01
	2620	8	0.03
	2630	6	0.02
	2640	1	0.00
	2650	3	0.01
	2660	6	0.02
	2670	3	0.01
	2680	1	0.00
	2690	4	0.01
	270	77	0.25
	2700	5	0.02
	2710	2	0.01
	2720	2	0.01
	2730	3	0.01
	2740	2	0.01
	2750	3	0.01
	2760	3	0.01
	2770	4	0.01
	2780	2	0.01
	280	59	0.19
	2800	3	0.01
	2810	3	0.01
	2820	3	0.01
	2830	5	0.02
	2840	4	0.01
	2850	3	0.01
	2860	4	0.01
	2870	2	0.01
	2880	2	0.01
	2890	3	0.01
	290	52	0.17

	2900	2	0.01
	2910	3	0.01
	2920	3	0.01
	2930	1	0.00
	2940	2	0.01
	2950	2	0.01
	2960	2	0.01
	2970	2	0.01
	2980	2	0.01
	2990	2	0.01
	30	1118	3.64
	300	599	1.95
	3000	4	0.01
	3010	1	0.00
	3020	2	0.01
	3030	2	0.01
	3040	2	0.01
	3050	8	0.03
	3060	1	0.00
	3070	2	0.01
	3080	4	0.01
	3090	2	0.01
	310	49	0.16
	3100	2	0.01
	3110	2	0.01
	3120	3	0.01
	3130	5	0.02
	3140	1	0.00
	3150	2	0.01
	3160	2	0.01
	3170	3	0.01
	3180	2	0.01
	3190	4	0.01
	32	4	0.01
	320	38	0.12

	3200	2	0.01
	3210	2	0.01
	3220	2	0.01
	3230	2	0.01
	3240	1	0.00
	3250	2	0.01
	3260	2	0.01
	3270	2	0.01
	3280	1	0.00
	3290	7	0.02
	330	226	0.74
	3300	1	0.00
	3310	2	0.01
	3320	2	0.01
	3330	2	0.01
	3340	2	0.01
	3350	2	0.01
	3360	2	0.01
	3370	2	0.01
	3380	2	0.01
	3390	1	0.00
	340	61	0.20
	3400	2	0.01
	3410	2	0.01
	3420	2	0.01
	3430	4	0.01
	3440	1	0.00
	3450	2	0.01
	3460	2	0.01
	3470	2	0.01
	3480	2	0.01
	3490	1	0.00
	350	43	0.14
	3500	1	0.00
	3510	2	0.01

	3530	2	0.01
	3540	2	0.01
	3550	1	0.00
	3570	1	0.00
	3580	1	0.00
	3590	1	0.00
	360	116	0.38
	3600	1	0.00
	3610	1	0.00
	3620	3	0.01
	3630	1	0.00
	3640	1	0.00
	3650	1	0.00
	3670	1	0.00
	3680	1	0.00
	3690	4	0.01
	370	62	0.20
	3700	1	0.00
	3710	2	0.01
	3720	1	0.00
	3730	1	0.00
	3740	1	0.00
	3750	1	0.00
	3760	1	0.00
	3770	1	0.00
	3780	1	0.00
	3790	1	0.00
	380	50	0.16
	3800	1	0.00
	3810	1	0.00
	3820	1	0.00
	3830	1	0.00
	3850	1	0.00
	3860	1	0.00
	3870	1	0.00

	3880	2	0.01
	3890	1	0.00
	390	66	0.21
	3900	1	0.00
	3910	3	0.01
	3920	1	0.00
	3930	1	0.00
	3940	1	0.00
	3950	1	0.00
	3960	2	0.01
	3970	1	0.00
	3980	1	0.00
	40	823	2.68
	400	32	0.10
	4000	1	0.00
	4020	1	0.00
	4030	2	0.01
	4040	1	0.00
	4050	1	0.00
	4060	1	0.00
	4070	2	0.01
	41	2	0.01
	410	56	0.18
	4100	3	0.01
	4110	1	0.00
	4120	1	0.00
	4130	1	0.00
	4140	1	0.00
	4150	3	0.01
	4160	1	0.00
	4170	1	0.00
	4180	1	0.00
	420	139	0.45
	430	38	0.12
	44	5	0.02

	440	42	0.14
	450	76	0.25
	460	48	0.16
	470	31	0.10
	480	26	0.08
	490	35	0.11
	50	564	1.83
	500	26	0.08
	510	38	0.12
	520	23	0.07
	530	28	0.09
	540	32	0.10
	550	24	0.08
	560	37	0.12
	570	25	0.08
	580	30	0.10
	590	25	0.08
	60	777	2.53
	600	27	0.09
	610	20	0.07
	620	189	0.61
	630	22	0.07
	640	27	0.09
	650	44	0.14
	660	32	0.10
	670	23	0.07
	680	30	0.10
	690	25	0.08
	70	487	1.58
	700	23	0.07
	710	24	0.08
	720	714	2.32
	730	27	0.09
	740	53	0.17
	750	16	0.05

	760	17	0.06
	770	24	0.08
	7700	3	0.01
	780	19	0.06
	790	57	0.19
	80	489	1.59
	800	35	0.11
	810	20	0.07
	820	18	0.06
	830	22	0.07
	840	17	0.06
	850	38	0.12
	860	15	0.05
	870	21	0.07
	880	77	0.25
	890	41	0.13
	90	344	1.12
	900	12	0.04
	910	9	0.03
	9100	2	0.01
	920	59	0.19
	9200	1	0.00
	930	19	0.06
	940	15	0.05
	950	12	0.04
	960	19	0.06
	970	72	0.23
	980	26	0.08
	990	18	0.06
BIOLIP	1	1098	4.18
	2	415	1.58
	3	447	1.70
	4	471	1.79
	5	340	1.30
	6	382	1.46

	7	252	0.96
	8	994	3.79
	9	1009	3.84
	10	453	1.73
	11	735	2.80
	12	581	2.21
	13	495	1.89
	14	490	1.87
	15	509	1.94
	16	973	3.71
	17	967	3.68
	18	999	3.81
	19	366	1.39
	20	330	1.26
	21	626	2.39
	22	509	1.94
	23	444	1.69
	24	485	1.85
	25	306	1.17
	26	171	0.65
	27	453	1.73
	28	243	0.93
	29	110	0.42
	30	344	1.31
	31	146	0.56
	32	88	0.34
	33	198	0.75
	34	187	0.71
	35	169	0.64
	36	221	0.84
	37	805	3.07
	38	150	0.57
	39	672	2.56
	40	216	0.82
	41	146	0.56

	42	194	0.74
	43	151	0.58
	44	53	0.20
	45	254	0.97
	46	374	1.43
	47	113	0.43
	48	575	2.19
	49	236	0.90
	50	64	0.24
	51	33	0.13
	52	176	0.67
	53	31	0.12
	54	58	0.22
	55	92	0.35
	56	66	0.25
	57	76	0.29
	58	56	0.21
	59	108	0.41
	60	95	0.36
	61	34	0.13
	62	126	0.48
	63	57	0.22
	64	76	0.29
	65	69	0.26
	66	21	0.08
	67	13	0.05
	68	38	0.14
	69	63	0.24
	70	85	0.32
	71	47	0.18
	72	83	0.32
	73	109	0.42
	74	124	0.47
	75	23	0.09
	76	3	0.01

	77	26	0.10
	78	94	0.36
	79	30	0.11
	80	16	0.06
	81	27	0.10
	82	30	0.11
	83	29	0.11
	84	50	0.19
	85	40	0.15
	86	41	0.16
	87	12	0.05
	88	117	0.45
	89	29	0.11
	90	39	0.15
	91	119	0.45
	92	85	0.32
	93	25	0.10
	94	18	0.07
	95	93	0.35
	96	20	0.08
	97	75	0.29
	98	75	0.29
	99	41	0.16
	100	57	0.22
	101	34	0.13
	102	8	0.03
	103	26	0.10
	104	23	0.09
	105	7	0.03
	107	15	0.06
	108	6	0.02
	109	7	0.03
	110	4	0.02
	111	4	0.02
	112	231	0.88

	113	37	0.14
	114	22	0.08
	115	3	0.01
	117	24	0.09
	118	6	0.02
	119	1	0.00
	122	2	0.01
	123	10	0.04
	125	22	0.08
	126	1	0.00
	127	33	0.13
	128	12	0.05
	129	28	0.11
	130	12	0.05
	131	4	0.02
	132	9	0.03
	133	15	0.06
	135	37	0.14
	136	6	0.02
	137	17	0.06
	138	7	0.03
	139	13	0.05
	141	4	0.02
	143	10	0.04
	144	10	0.04
	145	26	0.10
	146	96	0.37
	147	7	0.03
	148	61	0.23
	149	3	0.01
	150	4	0.02
	151	26	0.10
	152	3	0.01
	153	68	0.26
	154	6	0.02

	156	3	0.01
	157	5	0.02
	158	24	0.09
	159	7	0.03
	161	4	0.02
	163	1	0.00
	164	2	0.01
	168	14	0.05
	169	14	0.05
	173	5	0.02
	174	1	0.00
	175	19	0.07
	177	1	0.00
	178	7	0.03
	179	4	0.02
	180	12	0.05
	182	5	0.02
	184	10	0.04
	186	23	0.09
	188	5	0.02
	192	4	0.02
	193	10	0.04
	195	11	0.04
	199	1	0.00
	202	2	0.01
	203	3	0.01
	204	41	0.16
	205	54	0.21
	207	3	0.01
	209	2	0.01
	211	7	0.03
	213	4	0.02
	217	14	0.05
	218	3	0.01
	219	13	0.05

	221	5	0.02
	222	1	0.00
	227	1	0.00
	228	4	0.02
	236	13	0.05
	237	1	0.00
	242	1	0.00
	247	4	0.02
	248	1	0.00
	251	1	0.00
	252	4	0.02
	253	12	0.05
	255	8	0.03
	262	5	0.02
	263	4	0.02
	267	48	0.18
	268	6	0.02
	271	4	0.02
	272	6	0.02
	274	2	0.01
	275	6	0.02
	276	2	0.01
	289	6	0.02
	290	3	0.01
	-	787	3.00