UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Bruno Rodrigues da Silveira

Aprendizado de máquina para prospecção de exosítios de proteínas como moduladores de interação proteína-proteína e suas interfaces com *hot spots* identificados

São José do Rio Preto
2024

Bruno Rodrigues da Silveira

Aprendizado de máquina para prospecção de exosítios de proteínas como moduladores de interação proteína-proteína e suas interfaces com *hot spots* identificados

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Conselho de Curso de Bacharelado em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Financiadora: FAPESP - Proc. número 23/13399-0

Orientador:

Prof. Dr. Geraldo Francisco Donegá Zafalon

São José do Rio Preto
2024

S587a

Silveira, Bruno Rodrigues
Aprendizado de máquina para prospecção de exosítios de proteínas como moduladores de interação proteína-proteína e suas interfaces com hot spots identificados / Bruno Rodrigues Silveira. -- São José do Rio Preto, 2024
43 p.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto
Orientador: Geraldo Francisco Donegá Zafalon

1. Ciência da computação. 2. Bioinformática. 3. Aprendizado do computador. 4. Biologia Computacional. 5. Interação proteína-proteína. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Bruno Rodrigues da Silveira

Aprendizado de máquina para prospecção de exosítios de proteínas como moduladores de interação proteína-proteína e suas interfaces com *hot spots* identificados

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Conselho de Curso de Bacharelado em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Financiadora: FAPESP

Comissão Examinadora:

Prof. Dr. Geraldo Francisco Donegá Zafalon
UNESP – Câmpus de São José do Rio Preto
Orientador

Prof. Dr. Carlos Roberto Valêncio
UNESP – Câmpus de São José do Rio Preto

Profa. Dra. Renata Spolon Lobato
UNESP – Câmpus de São José do Rio Preto

São José do Rio Preto
2024

Agradecimentos

Agradeço primeiramente à minha familia por tudo que fizeram por mim até aqui e aos meus amigos de curso que sempre estiveram ao meu lado e me auxiliaram nas minhas situações dificeis durante esses quatro anos de faculdade.

Agradeço meus orientadores e também amigos, Geraldo e Vitória, que me auxiliaram além do ambiente escolar. Agradeço pelo carinho e por acreditarem no meu potencial.

Agradeço, por fim, à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo apoio financeiro a este trabalho, concedido por meio do processo 2023/13399-0.

“Be safe, friend. Don’t you dare go Hollow.”

- Laurentius of the Great Swamp (Dark Souls)

Resumo

Interações proteína-proteína e proteína-peptídeos ocorrem devido a estruturas presentes na camada superficial de proteínas chamadas de exosítios. Dentro das interfaces de ligações existem pequenas regiões que contribuem mais do que outras para a afinidade e especificidade da ligação, os chamados *hot spots*. Entender como essas interações ocorrem e são reguladas permitiria *insights* sobre os mecanismos de doença para o desenvolvimento de novas drogas e vacinas. Entretanto, descobrir novos tipos dessas interfaces é um trabalho manualmente trabalhoso e custoso, o que por muitas vezes inviabiliza a identificação e análise dos mesmos. Dessa forma, o presente projeto propõe a utilização de técnicas de *machine learning ensemble* para o desenvolvimento de um classificador capaz de identificar, por meio de diversos descritores físico-químicos, regiões características de exosítios e suas interfaces com *hot spots* identificados.

Palavras-chave: Ciência da computação; Bioinformática; Aprendizado do computador; Biologia Computacional; Interação proteína-proteína

Abstract

Protein-protein and protein-peptide interactions occur due to structures present on the protein surface known as exosites. Within the binding interfaces, there are small regions that contribute more than others to the binding affinity and specificity, called hot spots. Understanding how these interactions occur and are regulated could provide insights into disease mechanisms for the development of new drugs and vaccines. However, discovering new types of these interfaces is a manually intensive and costly task, often making their identification and analysis unfeasible. Therefore, the present project proposes the use of ensemble machine learning techniques to develop a classifier capable of identifying, through various physicochemical descriptors, characteristic regions of exosites and their interfaces with identified hot spots.

Keywords: Computer Science; Bioinformatics; Machine Learning; Computational Biology; Protein-Protein-Interface

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AF2 - *AlphaFold2*

CNN - *Convolutional Neural Network*

DPIs - *Drug–Protein Interactions*

EFB - *Exclusive Feature Bundling*

GBDT - *Gradient boosted decision tree*

GOSS - *Gradient-Based One-Side Sampling*

KS - *Kolmogorov-Smirnov*

ML - *Machine Learning*

MRMR - *Maximum Relevance Minimum Redundancy*

PCA - *Principal Component Analysis*

PPI - *Protein-Protein Interface*

PSSM - *Position-Specific Scoring Matrix*

SVM - *Support vector machine*

Sumário

Capítulo 1

Introdução ao trabalho

1.1 Introdução

As interações proteína-proteína causam uma ampla gama de processos biológicos, incluindo interações célula-a-célula e controle metabólico e de desenvolvimento (BRAUN; GINGRAS, 2012). Esse campo de estudo está se tornando um dos principais objetivos da biologia de sistemas, visto que contatos entre as cadeias de resíduos são a base para o dobramento (*folding*), montagem (*docking*) e interações entre proteínas (OFRAN; ROST, 2003), que induzem uma variedade de associações (BRAUN; GINGRAS, 2012; OFRAN; ROST, 2003; RAO et al., 2014). Assim, descobrir informações sobre interações proteína-proteína ajuda na identificação de alvos para medicamentos (RAO et al., 2014; PEDAMALLU; POSFAI, 2010).

Para muitas enzimas, a especificidade do substrato é dirigida por sítios de ligação secundários, chamados exosítios, que se diferem dos sítios atívos (YEGNESWARAN et al., 2007). Devido a isso, os exosítios tornaram-se tema de crescente interesse na pesquisa biomédica como potenciais alvos para medicamentos (YEGNESWARAN et al., 2007).

Para auxiliar no processo de entendimento e predição do comportamento de proteínas, técnicas de *machine learning* (ML) vêm sendo aplicadas com sucesso, acelerando

a compreensão sobre os aspectos intrínsecos às proteínas, como suas estruturas e funções (MORGUNOV et al., 2021). Ao identificar padrões presentes nos subconjuntos de dados detentores de um mesmo rótulo, métodos supervisionados de ML passam a procurá-los em outros conjuntos, sendo seu objetivo a definição de um rótulo, dentre os já conhecidos, para a informação de entrada (RUSSELL; NORVIG, 2016)

1.2 Objetivos

O objetivo do presente projeto é o desenvolvimento de um novo classificador utilizando técnicas de *machine learning* e *ensemble learning* para estudar a interação dos *hot spots* com interfaces proteína-proteína (PPI), a fim de encontrar novas possíveis interações com aminoácidos que possam ser utilizadas em fármacos. Dessa forma, a contribuição científica do projeto se mostra no processo inédito de classificação de *Hot spots* no substrato das proteínas.

1.3 Justificativa

Famílias de enzimas costumam compartilhar estruturas semelhantes e, dentro de cada família, seus sítios catalíticos são altamente conservados, de modo que um mesmo inibidor de enzimas possa inibir não só a sua enzima alvo, como também as enzimas homólogas a ela que tenham efeitos secundários indesejados (RIBEIRO et al., 2010). A identificação de exosítios permite que suas respectivas enzimas sejam direcionadas e inibidas por meio destes (BRUNGER; JIN; BREIDENBACH, 2008).

Diferentes regiões da superfície das proteínas apresentam diferentes características físico-químicas. Dessa forma, algumas regiões podem ser fragmentadas em regiões hidrofóbicas, denominadas de *hot-spots*. Estas regiões por não estabelecerem ligações com moléculas de água, são predominante locais de alta energia de contatos entre a proteína e seu ligante e proteína-proteína (SALIM, 2015), o que faz com que eles

possam ser a base estrutural necessária durante o desenho de drogas.

No que diz respeito ao uso de *machine learning*, um dos pilares da biologia computacional estrutural, este se justifica dada sua ampla aplicação em contextos médicos e biológicos. O aprendizado supervisionado, em especial, se destaca por seu excelente desempenho ao lidar com dados complexos e por se beneficiar da quantidade e diversidade de dados biológicos obtidos a partir do advento de sequenciadores automáticos (JIANG; GRADUS; ROSELLINI, 2020).

1.4 Motivação

A maioria dos processos moleculares e celulares é controlada por interações proteína-proteína que ocorrem através de interfaces presentes na superfície da proteína. A distribuição de energia ao longo da região de interface não é homogênea, certos resíduos contribuem mais para a energia livre de ligação, chamados de *hot spots* (TUNCBAG; KESKIN; GURSOY, 2010). Assim, aplicar métodos de aprendizado de máquina para identificar exosítios como moduladores das interações proteína-proteína com *hot spots* é uma abordagem favorável para a descoberta de novas ligações.

1.5 Metodologia

Inicialmente foram feitas pesquisas acerca da estrutura das proteínas a fim de entender o material de estudo. Alguns temas principais foram: exosítios, *hot spots*, sítios ativos, e interfaces proteína-proteína.

Em seguida, estudos sobre métodos de filtragem de características foram conduzidos para tratar os descritores das bases de dados STING_RDB. As *features* inicialmente foram selecionadas com base no trabalho de Pereira (2012), visto que o autor dedicou sua pesquisa para entender as importância de resíduos para as interações proteína-proteína e para predizer *hot spots*, alinhando com o objetivo do atual trabalho.

Em seguida, foi aplicado uma seleção baseada no teste de Kolmogorov-Smirnov (KS). Esse método busca avaliar duas características de mesma natureza por meio de uma função distribuição empírica seguindo a regra:

$$D_{x,y} = sup \forall z \in (-\inf, \inf) |P(X \leq z) - P(Y \leq z)|$$

Onde *sup* representa o valor supremo da operação e P é uma função distribuição. O objetivo de aplicar esse método é verificar se o descritor segue a hipótese nula de que ambos tem uma distribuição idêntica, tornando-os assim irrelevante para o resultado (IVANOV; RICCARDI, 2012). Por fim, foi feita uma análise individual de instâncias redundantes e faltantes para evitar ruídos e valores enviesados que poderiam atrapalhar o treinamento. Observou-se que muitos descritores traziam redundâncias nos seus dados, então, aplicou-se metodos para reduzí-los a fim de evitar *overfitting*, como por exemplo a função *drop_duplicates* que remove instâncias duplicadas e a *interpolate* que atribui um valor intermediario as instâncias faltantes, ambas da biblioteca Pandas. Além disso, pela complexidade da base, as classes apresentaram desbalanceamento após a aplicação do rótulo que foi tratado por meio do método *Random Under Sampler* que reduz o número de instâncias da classe predominante a fim de aproxima-las.

Alguns métodos de aprendizado de máquina foram selecionados para modelagem do classificador com base na sua qualidade de abstração dos resultados e com base em trabalhos correlatos onde resultados satisfatórios foram obtidos ao trabalhar com dados do STING_RDB. Os algoritmos foram *random forest*, *XGbooster* e *lightGBM*. Esses métodos foram implementados para testes a fim de escolher um deles para servir de base para o classificador.

Para avaliar o modelo, algumas métricas foram utilizadas como Acurácia, Precisão, *Recall* e *F1 Score*, utilizadas da biblioteca *scikit-learn*. A acúracia foi utilizada para medir a proporção de previsões corretas sobre o total de valores de teste. Seu intúito foi avaliar o desempenho global do modelo. A precisão foi utilizada para avaliar

proporção de previsões corretas em relação a outras predições de uma classe específica. Este método é importante para avaliar a capacidade classificativa do modelo. O *recall* foi utilizado para avaliar proporção de exemplos verdadeiramente positivos. Essa métrica é importante para verificar o potencial preditivo de cada classificador. O *F1 score* avaliou a média harmônica entre a precisão e o *recall*. Este foi implementado para medir o equilibrio entre essas duas métricas.

Além dos dados da STING_RDB, foi utilizada também a base de dados *PPI-HotSpot* da *Academia Sinica*. Diferente da STING_RDB, essa base possui informações apenas sobre *hot spots*, permitindo uma análise minusciosa sobre quais características buscar ao estudar as interações envolvendo essas regiões.

Com relação aos materiais, em termos de hardware, foram utilizadas as máquinas do Laboratório de Bioinformática da UNESP de São José do Rio Preto. Em termos de software, o ambiente Vim foi utilizado por meio de uma chave SSH com o servidor da Embrapa para execução e teste dos algoritmos que foram implementados na linguagem Python, devido a sua vasta disponibilidade de bibliotecas. Além disso, o aplicativo *putty* foi utilizado para acessar o banco de dados com os descritores.

Capítulo 2

Revisão Bibliográfica

2.1 Fundamentação Teórica

2.1.1 Proteínas

Proteínas são estruturas formadas pela combinação de 22 diferentes tipos de aminoácidos se difereciando em diversos aspectos, tais como tamanho das cadeias peptídicas, sequências dos aminoácidos utilizados, conformações tridimensionais, propriedades químicas dos aminoácidos que as compõem, entre outros (SALIM, 2015). Um dos principais grupos formados por interações entre proteínas são as enzimas. Enzimas agem como catalisadores para as diversas reações bioquímicas do corpo pois essas possuem um sítio ativo que é especificamente projetado para ligar apenas ao substrato pretendido da enzima (YEGNESWARAN et al., 2007).

Interações Proteína-Proteína

As interações proteína-proteína desempenham papéis cruciais na regulação de processos biológicos. Seus sítios de ligação são chamados de interfaces e as propriedades dos resíduos nas interfaces são os elementos-chave no reconhecimento, ligação e afinidade entre proteínas. Alterações nas interações proteína-proteína podem levar a várias

doenças, portanto, estudar as interfaces entre proteínas tem um enorme potencial na descoberta de medicamentos (CUKUROGLU et al., 2014).

Sítios Ativos

Um sítio ativo é uma pequena região da proteína onde ocorrerá a reação catalisada pela enzima. Esta é, portanto, a região da enzima que contém os resíduos de aminoácidos capazes de interagir com o substrato. É nesse sítio, também, que estão localizados os resíduos de aminoácidos que diretamente participam da ruptura e estabelecimento de ligações químicas que resultam na formação do produto (SALIM, 2015). Além desses sítios, existem sítios de ligações secundários que garantem a especificidade do subtrato e caracterizam as interações, os chamados exosítios.

Exosítios

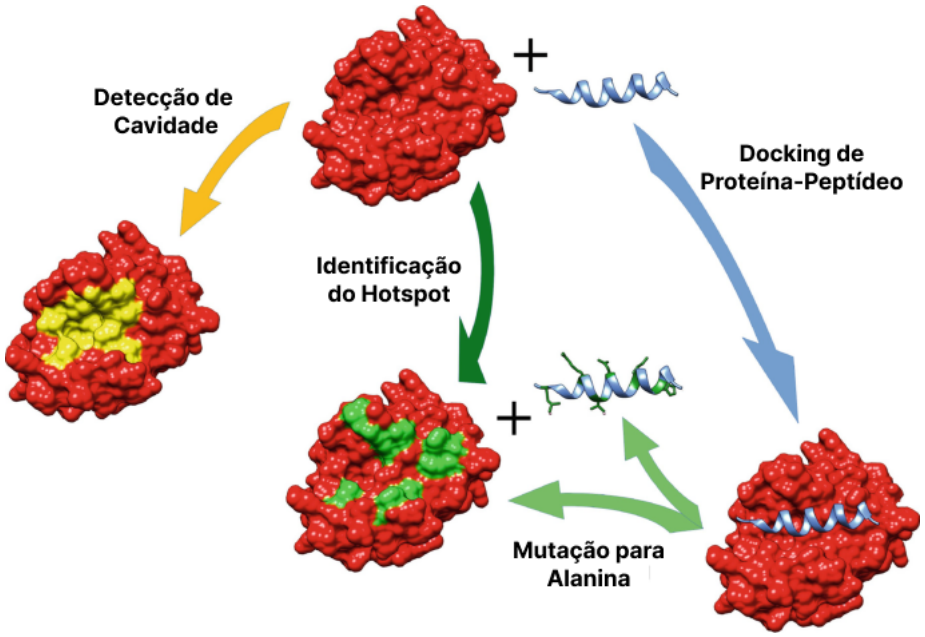
Os exosítios são estruturas presentes na camada superficial de proteínas que permitem que interações proteína-proteína e proteína-peptídeos ocorram e podem ser identificados por meio da análise de seus descritores físico-químicos e estruturais (YEGNESWARAN et al., 2007). O reconhecimento de uma macromolécula se dá por uma interação com os exosítios, que determinará a afinidade entre a enzima e o subtrato (KRISHNASWAMY; BETZ, 1997). Devido a isso, os exosítios tornaram-se tema de crescente interesse na pesquisa biomédica como potenciais alvos para medicamentos (YEGNESWARAN et al., 2007). No atual projeto, após a identificação dos exosítios, será possível modular novas possíveis interações proteína-proteína e modular *hot spots* nos resíduos da interface.

Hot Spots

Os resíduos nas interfaces proteína-proteína não contribuem igualmente para as energias de ligação (CUKUROGLU et al., 2014). Alternativamente, existem re-

giões de foco de energia de ligação chamados de *hot spots*. Um resíduo de interface é definido como resíduo de *hot spot* se sua mutação para alanina produzir um $\Delta\Delta G >= 2{,}0 kcal/mol$ (TUNCBAG; GURSOY; KESKIN, 2009), tornando-o mais suscetível a interações em interfaces proteína-proteína (BOGAN; THORN, 1998), o que leva ao ponto chave dessa pesquisa. É possível visualizar na figura 2.1 a detecção de um *hot spot* por meio do processo mencionado.

Figura 2.1: Método de detecção de um *hot spot*



Fonte: Adaptado de Delaunay e Ha-Duong (2022)

2.1.2 Aprendizado de Máquina

Aprendizado de máquina ou *machine learning* é o estudo de algoritmos e modelos estatísticos aplicados a sistemas computacionais para realização uma tarefa específica sem que sejam explicitamente programados para isso (MAHESH, 2020).

Algoritmos de aprendizado de máquina atuam das mais diversas formas e podem ser divididos de acordo com o método de aprendizagem utilizado (SALIM, 2015). Alguns métodos de aprendizado de máquina utilizados neste trabalho são: Aprendizado supervisionado e *ensemble learning*.

No aprendizado supervisionado, toda entrada é acompanhada de uma saída desejada que o método deve ser capaz de reproduzir. Assim, dado um conjunto de treinamento, o algoritmo busca encontrar uma função com menor erro na saída, visando homogeneizar a capacidade de generalização do modelo (SALIM, 2015). Para esse projeto, os dados de entrada foram os descritores de interfaces proteína-proteína, fornecidos pelo banco de dados STING_RDB e a saída será a prospecção dos exosítios. Além disso, terão dados de entradas referentes aos resíduos dos *hot spots* e não *hot spots*, que deverá resultar em uma saída classificando entre essas possíbilidades.

Ensemble learning é o processo pelo qual múltiplos modelos são gerados e combinados para resolver um determinado problema de inteligência computacional. O aprendizado *Ensemble* é utilizado principalmente para melhorar o desempenho de um classificador ou reduzir a probabilidade de uma seleção ruim de um modelo (MAHESH, 2020). Como citado na sessão 1.5, os modelos selecionados para testes com a base STING_RDB a fim de verificar qual terá os resultados mais adequados para o problema são o *random forest*, o *XGbooster* e o *lightGBM*.

2.1.3 Algoritmos

Random Forest

Random forest é um método *ensemble* baseado em *decision trees*. O algoritmo seleciona aleatoriamente características de entrada para gerar árvores de decisões individuais. A força de cada árvore de decisão determina o erro de generalização do classificador (BREIMAN, 2001; KULKARNI; SINHA, 2012).

XGboost

O algoritmo XGBoost é um método *ensemble* baseado em *gradient boosted decision tree* (GBDT), que é baseado em árvores de decisão. Seu ponto forte é ser escalável e preciso quando usado para aplicações de classificação e regressão. Diferentemente

do *gradient boosting*, esse possui uma função de regularização que evita o *overfitting* (MIENYE; SUN, 2022).

LightGBM

LightGBM é um algoritmo também baseado em GBDT (MIENYE; SUN, 2022), porém, o mesmo utiliza duas técnicas chamadas *Gradient-Based One-Side Sampling* (GOSS), que remove uma parcela de instâncias de dados com pequenos gradientes e usa apenas o restante para estimar o ganho de informação, e *Exclusive Feature Bundling* (EFB), que grupa características que raramente assumem valores diferentes de zero para reduzir o número de *features*, permitindo que o algoritmo seja executado mais rapidamente, mantendo um alto nível de precisão (KE et al., 2017).

2.1.4 Seleção de características

O processo de seleção de características, ou *feature selection*, é um método de redução de dimensionalidade no qual as características redundantes, irrelevantes ou que possam atrapalhar são removidas e um grupo de características consideradas relevantes é selecionado entre as *features* disponíveis de uma determinada base de dados (WANG; TANG; LIU, 2017). Para desenvolvimento do projeto, inicialmente as *features* de dados tabulares foram selecionadas da STING_RDB com base na sua significância biológica para o problema, visto que os exosítios podem ser identificados por meio da análise de seus descritores. Alguns métodos de *feature selection* são: *filters*, *wrappers* e *embeddeds*.

Nos métodos *filters*, os *features* são selecionadas com base apenas nas características dos dados, sem a utilização de um classificador. Neles, os atributos são avaliados individual e coletivamente e, com base nisso, são feitas análises estatísticas para definir quais são suas características mais relevantes (WANG; TANG; LIU, 2017; FURLANETTO; GOMES; BREVE, 2023). Para o atual projeto, o teste de *KS* foi a abordagem

utilizada sendo este um método *filter* baseado no modelo matemático de Kolmogorov-Smirnov para estudar a correlação entre os descritores (IVANOV; RICCARDI, 2012).

Modelos *wrapper* usam algoritmos de aprendizado específicos para avaliar a qualidade das *features* selecionadas. O componente de busca produzirá um conjunto de características com base em determinadas estratégias que podem estar inclusas na função objetiva do problema (WANG; TANG; LIU, 2017; FURLANETTO; GOMES; BREVE, 2023). Alguns exemplos desse método são o *Recursive Feature Elimination* e o *Sequential Feature Selection* que removem ou inserem características iterativamente (SAMB et al., 2012; KUNCHEVA, 2007).

Modelos *embedded* são baseados nos outros dois modelos, integrando a seleção de características na construção do modelo. Assim, os modelos *embedded* aproveitam tanto os modelos *filter* quanto os modelos *wrapper*. Seu ponto negativo é a não interpretabilidade dos resultados que pode levar à seleção sem sentido de características (WANG; TANG; LIU, 2017). O principal exemplo desse método é o *LASSO* que adiciona uma penalização na função de perda, orçando alguns coeficientes a se tornarem zero (RANSTAM; COOK, 2018).

2.1.5 STING_RDB

o STING_RDB é um banco de dados relacional de parâmetros estruturais para análise de proteínas com suporte para *data warehousing* e *data mining*. Este permite integração, aumento devido a escalabilidade juntamente do perfil de dados que permite consultar, monitorar e inspecionar suas instâncias por meio de consultas ao servidor. Além disso, o banco mantém qualidade dos dados devido a constante manutenção de valores que são atualizados em tempo real (OLIVEIRA et al., 2007).

Atualmente, o banco de dados STING_RDB possui mais de 300 descritores compilados em um único site. O mesmo foi implementado usando *MySQL*, um sistema de gerenciamento de banco de dados relacional de código aberto que usa linguagem de

consulta estruturada (SQL). Esse banco de dados está conectado ao STING, um conjunto de programas baseado na *web* para análise simultânea de estruturas e sequências (OLIVEIRA et al., 2007).

2.2 Trabalhos Correlatos e Estado da Arte

Trabalhos como o de Zhou et al. (2024) mostram a necessidade emergente de modelos de predição de interações entre proteínas. O mesmo busca em sua pesquisa elaborar um modelo usando pré-treinamento em larga escala, mecanismos de difusão global e extração de subgrafos locais para inferir novas interações entre proteínas e drogas (DPIs). Os resultados obtidos mostraram que o modelo demonstrou eficiência na inferência de novas potenciais interações, apesar de não se destacar ao buscar por interações já existentes. Assim, o autor conclui reforçando a necessidade de desenvolver um modelo eficiente para descobrir mais DPIs desconhecidos (ZHOU et al., 2024).

Já em Le et al. (2024) buscou-se estudar, além das iterações entre proteínas e drogas, sítios de interação proteína-peptídeo por meio de uma ferramenta chamada *ProtTrans*, um modelo de linguagem pré-treinado especificamente projetado para sequências de proteínas, juntamente de redes neurais convolucionais (CNN) de multi-janela. É mencionado no artigo a relevância dessas inferências de novas interações de proteínas, sendo a melhor compreensão para desenvolvimento de medicamentos antivirais ou vacinas para doenças infecciosas, como gripe, HIV ou COVID-19. Assim, usando o método de validação cruzada de 5 partes, foi realizada uma *cross-validation* no conjunto de dados de treinamento para avaliar o desempenho geral dentro dos critérios de sensibilidade, especificidade, precisão e acurácia. Com isso, o autor obteve resultados satisfatórios quando comparado a outros métodos de aprendizado de máquina o que leva a reforçar o bom desempenho de métodos *ensemble* (LE et al., 2024).

O trabalho de Bryant, Pozzati e Elofsson (2022) também busca predizer interações

entre proteínas. Entretanto, para este caso, os autores utilizaram a ferramenta *AlphaFold2* (AF2), uma ferramenta de predição de estrutura e modelagem tridimensional de proteínas (JUMPER et al., 2021), para predizer interações proteína-proteína. No caso, dois conjuntos de dados diferentes foram selecionados para serem modelados e ancorados simultaneamente a fim de estudar a relação entre a qualidade do modelo de saída e as entradas. Com base nisso, os autores conseguiram pontuar vários modelos de interação proteína-proteína com uma pontuação chamada *predicted DockQ score*, que mostra a relação entre as iterações, assim obtendo uma acurácia de 57,8%. Apesar dos resultados, o artigo cita que houveram limitações para a aplicação da ferramenta onde as amostras tiveram de ser de menor complexidade para o funcionamento dos testes (BRYANT; POZZATI; ELOFSSON, 2022).

De maneira similar ao presente trabalho, Qiao et al. (2018) utiliza *support vector machine* (SVM) como modelo de classificação com a função de base radial de *kernel*. A ideia proposta por esse trabalho é uma seleção de características hibrida, baseada nos resultados de árvores de decisão, *F-score* e *maximum relevance minimum redundancy* (mRMR), de forma que as *features* selecionados fossem considerados para treinamento do SVM. A base de dados utilizada inclui 154 resíduos de interface com diferenças de energia observadas em um método chamado varredura de alanina, que permitiu a classificação entre *hot spots* e não *hot spots*. Quanto aos resultados, o autor não obteve bons resultados de precisão, especificidade e exatidão quando avaliadas separadamente pois o mesmo diz que esses três parâmetros não devem ser avaliados individualmente, visto que pode levar a *overfitting*. Por outro lado, o autor também utilizou um critério *recall* que avalia apenas a precisão preditiva de exemplos positivos. Com isso, ainda sim foi possível avaliar biologicamente as características selecionadas de forma que houvesse significancia no resultado (QIAO et al., 2018).

Por sua vez, Moreira et al. (2017) buscaram identificar e classificar resíduos de interface como *hot spots* e não *hot spots* (*null spots*). Porém, neste caso, os autores

desenvolveram uma ferramenta *web* chamada *SpotOn*, utilizando uma abordagem *ensemble* treinada com 53 complexos e usando várias *features* baseadas tanto na estrutura 3D da proteína quanto na sequência. Também é mostrada a importância da qualidade das características selecionadas e bases de dados escolhidas, visto que foi um gargalo para esta pesquisa. O autor cita que os dados utilizados, como a maioria dos dados em biologia, são atípicos para aprendizado de máquina, sendo muito esparsos, incompletos, tendenciosos e ruidosos, sendo marcado por dados desbalanceados, o que tornou a seleção de medidas de desempenho e algoritmos apropriados ainda mais necessário para obter um bom resultado (MOREIRA et al., 2017).

Capítulo 3

Desenvolvimento do Trabalho

Conforme citado na Sessão 1.5, pesquisas acerca das estruturas e dos métodos foram realizados como etapa inicial do projeto com o objetivo de compreender seu funcionamento e assim auxiliar na elaboração da estratégia proposta neste trabalho. O tratamento da base, o pré processamento das características e a criação dos classificadorres serão detalhadas nas seguintes subseções.

3.1 Base STING_RDB

Referente aos dados utilizados para esse projeto, foi solicitado acesso ao banco de dados da Embrapa, o STING_RDB, que contém descritores físico-químicos utilizados para prospecção de exosítios. Com o acesso ao banco, alguns descritores foram selecionados com base no trabalho de Pereira (2012), visto que o autor dedicou sua pesquisa para melhor entender os descritores de proteínas e sua significância para os resíduos *hot spots*. Por meio de consultas em *MySQL*, as características selecionadas são apresentadas na Tabela 3.1.

Tabela 3.1: Tabela de descritores de proteínas

Descritores de proteínas	
Descritor	Descrição
pdb_id	Id da proteína no *PDB*
pdb_chain	Cadeia polipeptídica da proteína
pdb_resno	Número do resíduo
ep_ca	Potencial eletrostático calculado no C_α
ep_lha	Potencial eletrostático calculado no último átomo da cadeia lateral não hidrogênio
ep_absolute	Soma do potencial eletrostático de todos os átomos do resíduo
ep_average	Média do potencial eletrostático de todos os átomos do resíduo
ep_surface	Potencial eletrostático calculado na superfície do resíduo
hydro_radzicka_isol_surfv	Hidrofobicidade de uma superfície por Radzicka
hydro_kite_dolitte_isol_surfv	Hidrofobicidade de uma superfície por Kyte-Doolittle
hydro_radzicka_complex_naccess	Acessibilidade do solvente em estruturas complexas de proteínas por Radzicka
hydro_kite_dolitte_complex_naccess	Acessibilidade do solvente em estruturas complexas de proteínas por Kyte-Doolittle
ced_CA	Densidade de energia calculada com esfera centrada no C_α e raio variando de 3 a 7
ced_LHA	Densidade de energia calculada com esfera centrada no último átomo da cadeia lateral não hidrogênio e raio variando de 3 a 7
ced_CA_IFR	Densidade de energia na interface calculada com esfera centrada no C_α e raio variando de 3 a 7

Continuação na próxima página...

Continuação da Tabela 3.1	
Descritor	Descrição
ced_LHA_IFR	Densidade de energia na interface calculada com esfera centrada no último átomo da cadeia lateral não hidrogênio e raio variando de 3 a 7
entd_CA	Densidade de entropia calculada com esfera centrada no C_α e raio variando de 3 a 7
entd_LHA	Densidade de entropia na interface calculada com esfera centrada no último átomo da cadeia lateral não hidrogênio e raio variando de 3 a 7
entd_CA_IFR	Densidade de entropia na interface calculada com esfera centrada no C_α e raio variando de 3 a 7
entd_LHA_3_IFR	Densidade de entropia na interface calculada com esfera centrada no último átomo da cadeia lateral não hidrogênio e raio variando de 3 a 7
weighted_contact_number	Número de contatos ponderado para um resíduo
avg_weighted_contact_number_k	Média do número de contatos ponderado para um conjunto de resíduos, considerando uma vizinhança variando de 2 a 5
catalytic_profile_comb_score_k	Pontuação combinada de perfil catalítico para uma janela de tamanho variando de 2 a 5
class	Rótulo atribuido como *hot spot* ou não *hot spot*

Totalizando 61 descritores e o rótulo para classificação dos algoritmos *ensemble*, essas foram as características selecionadas para o pré processamento.

3.2 Pré processamento dos dados

Antes de aplicar os algoritmos de aprendizado, foi necessário tratar os dados pois alguns descritores estavam redundantes e algumas instâncias estavam incompletas. Como mencionado na seção 2.1.4, um dos métodos utilizados foi o teste KS para seleção de características a fim de encontrar *features* que possuem uma distribuição semelhante. A implementação dessa função foi feita por meio da biblioteca *scipy* e da função *ks_2samp* que compara dois descritores por meio da variável *p_value*, cujo critério foi eliminar descritores com uma semelhança maior do que 5% do *p_value*.

Por tratar-se de uma base com muitas instâncias, suas entradas possuiam valores redundantes e ruídos que atrapalhavam o aprendizado, tornando-o enviesado. Assim, utilizou-se alguns métodos da biblioteca Pandas como *interpolate* para atribuir valor a variáveis não numéricas e *drop_duplicates* para remover valores duplicados.

Por fim, após o tratamento dos dados, foi necessário balancear as classes, visto que *hot spot* é a menos predominante no *dataframe*. Para isso, foi utilizada a função *RandomUnderSampler* da biblioteca *imbalanced-learn*, cuja função foi diminuir a classe prevalente de forma que ambas tivessem tamanhos próximos sem perder a essência dos dados.

3.3 Criação do classificador

A escolha dos algoritmos para o atual projeto deu-se priorizando abordagens *ensemble*, visto que essas técnicas apresentaram melhores resultados em trabalhos relacionados da área. Dentre esses, os algoritmos *tree based* foram escolhidos devido à sua melhor capacidade de abstração dos resultados. Por se tratarem de dados biológicos, é importante entender a significância dos valores obtidos. Dessa forma, foram utilizados os algoritmos *Random Forest*, *LightGBM* e *XGBoost* através das bibliotecas *scikit-learn*, *lightgbm* e *xgboost*, respectivamente.

A divisão das bases de treinamento e de teste foi feitas por meio da função *train_test_split*, garantindo a separação válida na proporção 30% teste e 70% treinamento.

Capítulo 4

Testes e Resultados

4.1 Realização dos testes

Os testes foram realizados utilizando uma máquina virtual em ambiente disponibilizado pela Embrapa, conectado por meio de uma chave SSH. O acesso ao banco de dados foi feito por meio de requisições em SQL utilizando a biblioteca *mysql.connector*, também por meio do servidor da Embrapa. Além disso, a base *PPI-Hotspot* foi integrada ao ambiente de forma que fosse possível fazer requisições unindo ambas as bases *STING_RDB* e *PPI-Hotspot*.

Para o *dataframe* criado, foi aplicado o teste de *KS*, um método *filter* para *feature selection*. Em seguida, para lidar com o desbalanceamento, a função *Random under Sampler* foi aplicada de modo a reduzir a classe "não *hot spot*". Por fim, a base foi dividida utilizando o método *train_test_split* e o modelo foi treinado e testado para 100 iterações, de forma que fosse possível observar resultados para diversos cenários.

As métricas adotadas para a avaliação dos modelos foram acurácia, precisão, *recall* e *F1 Score*, utilizadas da biblioteca *scikit-learn*. As tabela 4.1, 4.2 e 4.3 mostram o desempenho de cada modelo para essas métricas

Além disso, na Figura 4.4 são apresentadas a média da relevância de cada *feature* a classificação dos modelos.

4.2 Resultados

Os resultados dos testes para os três algoritmos, suas comparações e a avaliação dos descritores geradas pelos modelos encontram-se nas subseções a seguir:

4.2.1 *Random Forest*

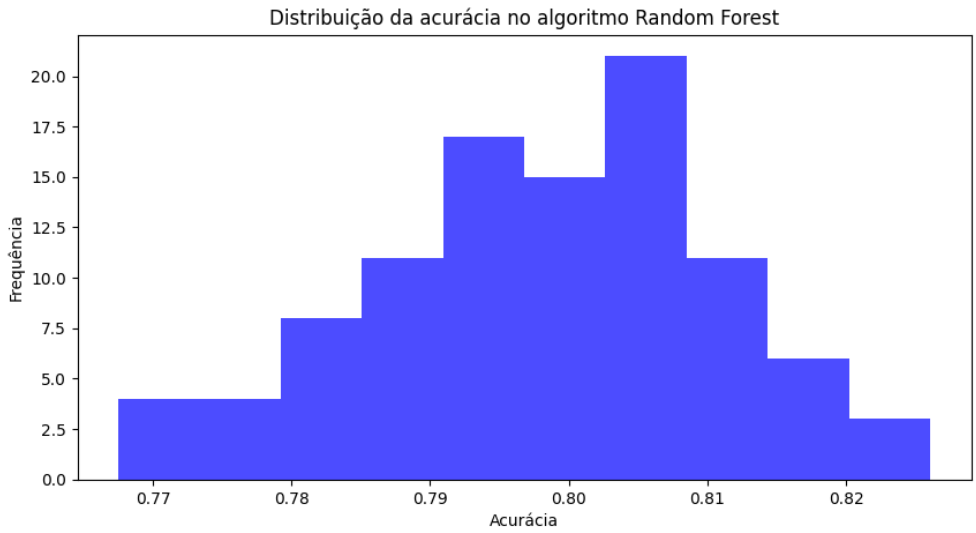
Como é apresentado na Tabela 4.1, o modelo *Random Forest* apresentou um desempenho consistente, com valores equilibrados entre precisão e recall, indicando que ele não está enviesado para uma classe específica. Quanto ao critério de acurácia, o modelo obteve bons resultados visto que as classes estavam balanceadas, indicando um bom desempenho para classificar o problema. Sua precisão e *recall* também foram boas, indicando que o modelo apresenta bom potencial classificativo e preditivo para o atual problema, que, por consequência, indica um bom *F1 Score*.

Tabela 4.1: Média dos resultados do modelo *Random Forest*

Random Forest	
Métrica	**Valor**
Acurácia	79,80%
Precisão	79,50%
Recall	80,33%
F1	79,89%

Na figura 4.1 apresenta-se oscilações no valor de acurácia do algoritmo. Devido a mudança na base de treinamento, reflexo da função *train_test_split*, e devido ao modelo *random forest* inicializar com árvores randomicas, sua acurácia oscilou entre as 100 iterações indicando resultados dispersos. Apesar disso, ainda sim o algoritmo variou entre 77% e 82%, mantendo um bom desempenho quanto aos acertos.

Figura 4.1: Distribuição da acurácia no modelo *Random Forest*



Fonte: Elaborado pelo próprio autor

4.2.2 *XGBoost*

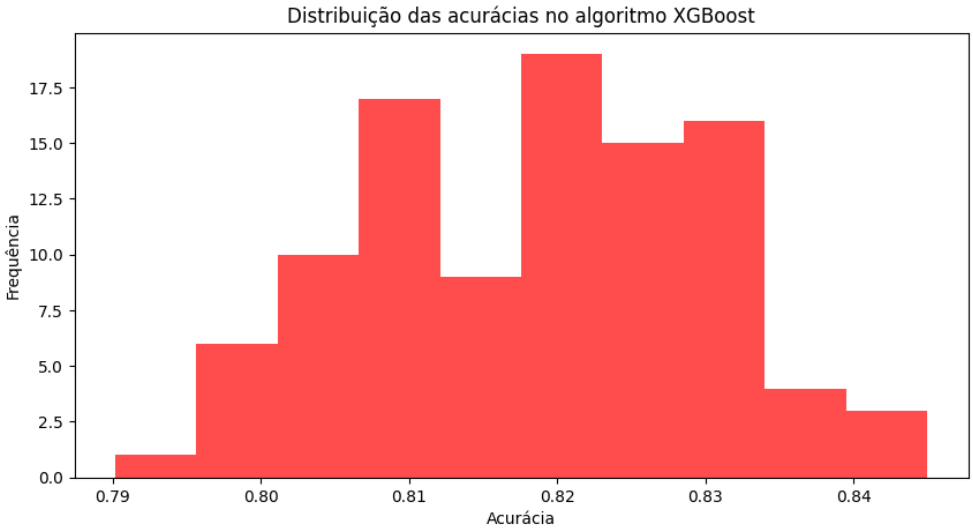
Assim como mostrado na Tabela 4.2, a precisão e o *recall* foram próximos e altos, refletindo assim sua acurácia alta e o bom desempenho global para esse modelo. O destaque desse algoritmo foi referente a sua alta média de *recall*, indicando que este modelo apresentou um alto poder preditivo, podendo ser muito útil para descobrir novos sítios de *hot spots* quando aplicado a casos preditivos. Por consequência dos demais parâmetros altos, seu *F1* refletiu isso também possuindo um resultado elevado.

Tabela 4.2: Média dos resultados do modelo *XGBoost*

XGBoost	
Métrica	**Valor**
Acurácia	81,85%
Precisão	81,12%
Recall	83,07%
F1	82,07%

É apresentado na Figura 4.2 o resultado do modelo *XGBoost* quando submetido as mesmas condições do algoritmo *random forest*, refletindo assim sua oscilação na acurácia. Seus valores variam entre 79% e 84%, tendo seu pico em 82%.

Figura 4.2: Distribuição da acurácia no modelo *XGBoost*



Fonte: Elaborado pelo próprio autor

4.2.3 *LightGBM*

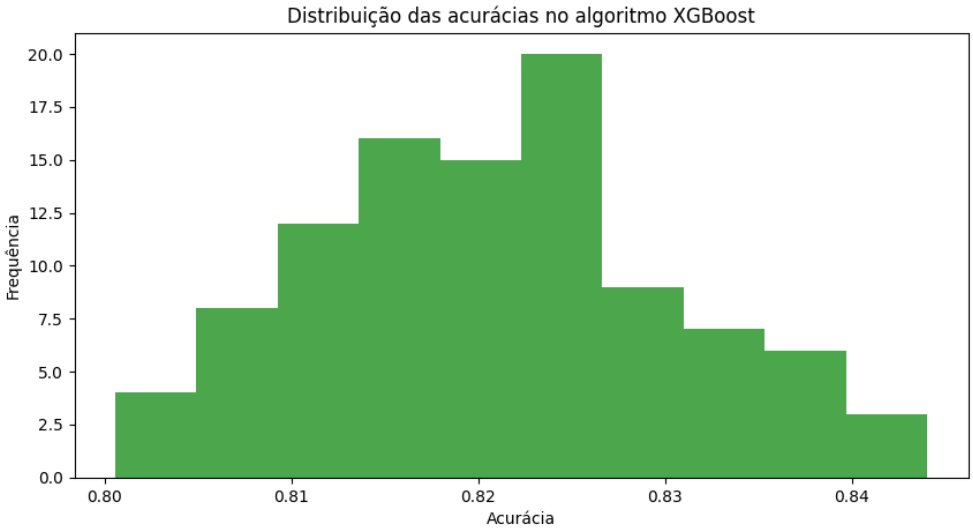
De acordo com a Tabela 4.3 é possível visualizar que todas as métricas obtiveram resultados satisfatórios para a predição. O algoritmo mostra-se eficaz em todas as métricas e faz previsões corretas com frequência, visto que a média de seus resultados obtiveram maiores pontos percentuais que os demais classificadores.

Tabela 4.3: Média dos resultados do modelo *LightGBM*

LightGBM	
Métrica	**Valor**
Acurácia	82,09%
Precisão	81,39%
Recall	83,33%
F1	82,33%

Dado que o modelo *LightGBM* também foi submetido as mesmas condições dos demais, a figura 4.3 mostra o intervalo de variação de sua acurácia. Os resultados do algoritmo oscilaram entre 80% e 84%, mostrando uma variação.

Figura 4.3: Distribuição da acurácia no modelo *LightGBM*



Fonte: Elaborado pelo próprio autor

4.2.4 Comparação entre os modelos

Os resultados obtidos para os três algoritmos *Random Forest*, *XGBoost* e *LightGBM* mostram diferenças sutis, mas importantes, que permitiram uma análise mais precisa sobre o desempenho de cada modelo como mostrado na tabela 4.4. Dentre eles, o *LightGBM* se mostrou a opção mais eficiente entre os critérios avaliados e em particular o *recall* que é um critério importante para possível predição de novos resídos de *hot spots*.

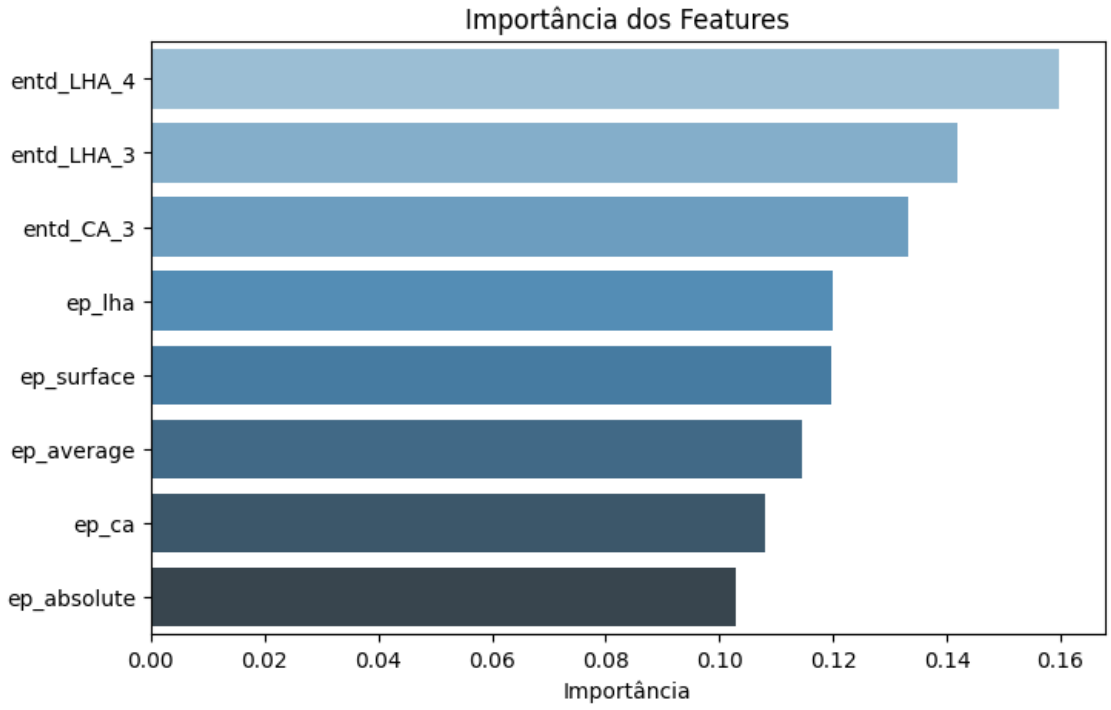
Tabela 4.4: Comparação entre os três algoritmos

Métrica	***Random Forest***	***XGBoost***	***LightGBM***
Acurácia	79,80%	81,85%	82,09%
Precisão	79,50%	81,12%	81,39%
Recall	80,33%	83,07%	83,33%
F1	79,89%	82,07%	82,33%

4.2.5 Avaliação dos descritores

Como apresentado na sessão 3.2, vários descritores foram eliminados ou por serem altamente correlacionados entre as classes ou por serem altamente redundantes. Após o treinamento e teste dos modelos, foi gerada a Figura 4.4, na qual é possível notar que o valor de relevância dessas *features* se concentraram entre a densidade de entropia e o potencial eletrostático, indicando que esses descritores podem estar diretamente relacionados com a detecção de resíduos de *hot spots* nas interfaces de ligações de proteínas. Devido a isso, é possível que, para trabalhos futuros, esses descritores tenham maior ênfase quando o foco de estudo forem os exosítios.

Figura 4.4: Relevância de cada descritor



Fonte: Elaborado pelo próprio autor.

Capítulo 5

Conclusão

Neste trabalho, foi apresentada uma análise de métodos *ensemble* aplicados à detecção de exosítios proteínas, em particular, *hot spots*, destacando o uso dos algoritmos *random forest*, *XGboost* e *LightGBM*. Assim como mostrado na sessão 2.2, estudos acerca dos sítios de ligações secundários por meio de técnicas aprendizado de máquina tem sido bastante eficientes para classificar e predizer essas regiões. A busca por novas ferramentas com esse poder preditivo é de suma importância para o *design* de drogas e desenvolvimento de novos fármacos (RAO et al., 2014; PEDAMALLU; POSFAI, 2010).

Os bons resultados do projeto se deram devido ao grande número de descritores disponibilizados pelo STING_RDB e a vasta gama de bibliotecas disponives ao utilizar ferramentas em python. Uma vez treinados, os algoritmos de *machine learning* poderão ser aplicados para casos mais complexos futuramente.

Dentre os descritores selecionados, 8 deles destacaram-se para o modelo de predição, obtendo mais de 98% de importância para classificar os resíduos da superfície. Com isso, foi possível notar a relevâncias das características de densidade de entropia e potencial eletrostático para predição das interfaces de *hot spots*.

Assim, o projeto buscou auxiliar o entendimento acerca das proteínas e de suas interações dentro de sistemas biológicos, como o corpo humano, com o intuito de

viabilizar o desenvolvimento de novas drogas, com foco na inibição de enzimas e também um maior controle de processos celulares mediados por proteínas alvo.

Referências Bibliográficas

BOGAN, A. A.; THORN, K. S. **Anatomy of hot spots in protein interfaces**. *Journal of molecular biology*, Elsevier, v. 280, n. 1, p. 1–9, 1998.

BRAUN, P.; GINGRAS, A.-C. **History of protein–protein interactions: From egg-white to complex networks**. *Proteomics*, Wiley Online Library, v. 12, n. 10, p. 1478–1498, 2012.

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

BRUNGER, A.; JIN, R.; BREIDENBACH, M. **Highly specific interactions between botulinum neurotoxins and synaptic vesicle proteins**. *Cellular and Molecular Life Sciences*, Springer, v. 65, p. 2296–2306, 2008.

BRYANT, P.; POZZATI, G.; ELOFSSON, A. ***Improved prediction of protein-protein interactions using AlphaFold2. Nat. Commun. 13, 1265***. 2022.

CUKUROGLU, E. et al. **Hot spots in protein–protein interfaces: Towards drug discovery**. *Progress in biophysics and molecular biology*, Elsevier, v. 116, n. 2-3, p. 165–173, 2014.

DELAUNAY, M.; HA-DUONG, T. **"Computational Tools and Strategies to Develop Peptide-Based Inhibitors of Protein-Protein Interactions"**. In: ____. *Computational Peptide Science: Methods and Protocols*. New York, NY: Springer US, 2022. p. 205–230. ISBN 978-1-0716-1855-4.

FURLANETTO, G. C.; GOMES, V. Z.; BREVE, F. A. **Artificial Bee Colony Algorithm for Feature Selection in Fraud Detection Process**. In: SPRINGER. *International Conference on Computational Science and Its Applications*. [S.l.], 2023. p. 535–549.

IVANOV, A.; RICCARDI, G. **Kolmogorov-Smirnov test for feature selection in emotion recognition from speech**. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2012. p. 5125–5128.

JIANG, T.; GRADUS, J. L.; ROSELLINI, A. J. **Supervised machine learning: a brief primer**. *Behavior therapy*, Elsevier, v. 51, n. 5, p. 675–687, 2020.

JUMPER, J. et al. **Highly accurate protein structure prediction with AlphaFold**. *nature*, Nature Publishing Group, v. 596, n. 7873, p. 583–589, 2021.

KE, G. et al. **Lightgbm: A highly efficient gradient boosting decision tree**. *Advances in neural information processing systems*, v. 30, 2017.

KRISHNASWAMY, S.; BETZ, A. **Exosites determine macromolecular substrate recognition by prothrombinase**. *Biochemistry*, ACS Publications, v. 36, n. 40, p. 12080–12086, 1997.

KULKARNI, V. Y.; SINHA, P. K. **Pruning of random forest classifiers: A survey and future directions**. In: IEEE. *2012 International Conference on Data Science & Engineering (ICDSE)*. [S.l.], 2012. p. 64–68.

KUNCHEVA, L. I. **A stability index for feature selection.** In: CITESEER. *Artificial intelligence and applications*. [S.l.], 2007. p. 421–427.

LE, V.-T. et al. **rotTrans and multi-window scanning convolutional neural networks for the prediction of protein-peptide interaction sites**. *Journal of Molecular Graphics and Modelling*, Elsevier, v. 130, p. 108777, 2024.

MAHESH, B. **Machine learning algorithms-a review**. *International Journal of Science and Research (IJSR).[Internet]*, v. 9, n. 1, p. 381–386, 2020.

MIENYE, I. D.; SUN, Y. **A survey of ensemble learning: Concepts, algorithms, applications, and prospects**. *IEEE Access*, IEEE, v. 10, p. 99129–99149, 2022.

MOREIRA, I. et al. ***Spoton: high accuracy identification of protein-protein interface hot-spots. Sci Rep 7 (1): 8007***. 2017.

MORGUNOV, A. S. et al. **New frontiers for machine learning in protein science**. *Journal of Molecular Biology*, Elsevier, v. 433, n. 20, p. 167232, 2021.

OFRAN, Y.; ROST, B. **Analysing six types of protein–protein interfaces**. *Journal of molecular biology*, Elsevier, v. 325, n. 2, p. 377–387, 2003.

OLIVEIRA, S. d. M. et al. **Sting_RDB: a relational database of structural parameters for protein analysis with support for data warehousing and data mining.** Genetics and Molecular Research, v. 6, n. 4, p. 911-922, 2007., 2007.

PEDAMALLU, C. S.; POSFAI, J. **Open source tool for prediction of genome wide protein-protein interaction network based on ortholog information**. *Source code for biology and medicine*, Springer, v. 5, p. 1–6, 2010.

PEREIRA, J. G. de C. ***Caracterização dos aminoácidos da interface proteína-proteína com maior contribuição na energia de ligação e sua predição a partir dos dados estruturais***. Dissertação (Mestrado) — Universidade Estadual de Campinas, Instituto de Biologia, Campinas, SP, 2012.

QIAO, Y. et al. **Protein-protein interface hot spots prediction based on a hybrid feature selection strategy**. *BMC bioinformatics*, Springer, v. 19, p. 1–16, 2018.

RANSTAM, J.; COOK, J. A. **LASSO regression**. *Journal of British Surgery*, Oxford University Press, v. 105, n. 10, p. 1348–1348, 2018.

RAO, V. S. et al. **Protein-protein interaction detection: methods and analysis**. *International journal of proteomics*, Wiley Online Library, v. 2014, n. 1, p. 147648, 2014.

RIBEIRO, C. et al. **Analysis of binding properties and specificity through identification of the interface forming residues (IFR) for serine proteases in silico docked to different inhibitors**. *BMC Structural Biology*, Springer, v. 10, p. 1–16, 2010.

RUSSELL, S. J.; NORVIG, P. ***Artificial intelligence: a modern approach***. [S.l.]: Pearson, 2016.

SALIM, J. A. **Aplicação de técnicas de reconhecimento de padrões usando os descritores estruturais de proteínas da base de dados do software STING para discriminação do sítio catalítico de enzimas**. *Master's thesis. Campinas: UNICAMP, Faculdade de Engenharia Elétrica e Computação*, 2015.

SAMB, M. L. et al. **A novel RFE-SVM-based feature selection approach for classification**. *International Journal of Advanced Science and Technology*, Citeseer, v. 43, n. 1, p. 27–36, 2012.

TUNCBAG, N.; GURSOY, A.; KESKIN, O. **Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy**. *Bioinformatics*, Oxford University Press, v. 25, n. 12, p. 1513–1520, 2009.

TUNCBAG, N.; KESKIN, O.; GURSOY, A. **HotPoint: hot spot prediction server for protein interfaces**. *Nucleic acids research*, Oxford University Press, v. 38, n. suppl_2, p. W402–W406, 2010.

WANG, S.; TANG, J.; LIU, H. ***Feature Selection.*** 2017.

YEGNESWARAN, S. et al. **Manipulation of thrombin exosite I, by ligand-directed covalent modification**. *Journal of Thrombosis and Haemostasis*, Elsevier, v. 5, n. 10, p. 2062–2069, 2007.

ZHOU, Z. et al. **Revisiting drug–protein interaction prediction: a novel global–local perspective**. *Bioinformatics*, Oxford University Press, v. 40, n. 5, p. btae271, 2024.