Algoritmo Genético Assistido por *Surrogate* para avaliar e descobrir peptídeos contra o SARS-CoV-2

Elias de Abreu Domingos da Silva

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2022

Elias de Abreu Domingos da Silva

Algoritmo Genético Assistido por *Surrogate* para avaliar e descobrir peptídeos contra o SARS-CoV-2

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Prof. Dr. Murillo Guimarães Carneiro
Coorientador: Prof. Dr. Luiz Gustavo Almeida Martins

Uberlândia
2022

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

S586 Silva, Elias de Abreu Domingos da, 1992-
2022 Algoritmo Genético Assistido por Surrogate para avaliar e descobrir peptídeos contra o SARS-CoV-2 [recurso eletrônico] / Elias de Abreu Domingos da Silva. - 2022.

Orientador: Murillo Guimarães Carneiro.
Coorientador: Luiz Gustavo Almeida Martins.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2022.571
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Carneiro, Murillo Guimarães,1988-, (Orient.). II. Martins, Luiz Gustavo Almeida,1974-, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Ciência da Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação de Mestrado 16/2022, PPGCO				
Data:	29 de setembro de 2022	Hora de início:	09:00	Hora de encerramento:	11:33
Matrícula do Discente:	12022CCP002				
Nome do Discente:	Elias de Abreu Domingos da Silva				
Título do Trabalho:	Algoritmo Genético Assistido por Surrogate para avaliar e descobrir peptídeos contra o SARS-CoV-2.				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Inteligência Artificial				
Projeto de Pesquisa de vinculação:	-				

Reuniu-se, por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Prof. Dr. Nilson Nicolau Júnior - IBTEC/UFU; Prof. Dr. Renato Tinós - USP; Prof. Dr. Luiz Gustavo Almeida Martins - FACOM/UFU e Prof. Dr. Murillo Guimarães Carneiro, orientador do candidato.

Os examinadores participaram desde as seguintes localidades: Renato Tinós - Ribeirão Preto/SP; Nilson Nicolau Júnior, Luiz Gustavo Almeida Martins e Murillo Guimarães Carneiro - Uberlândia/MG. O discente participou da cidade de Uberlândia/MG.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Murillo Guimarães Carneiro, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Renato Tinós**, **Usuário Externo**, em 04/10/2022, às 09:55, conforme horário oficial de Brasília, com fundamento no art. 6°, § 1°, do Decreto n° 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Luiz Gustavo Almeida Martins**, **Professor(a) do Magistério Superior**, em 04/10/2022, às 10:01, conforme horário oficial de Brasília, com fundamento no art. 6°, § 1°, do Decreto n° 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Nilson Nicolau Junior**, **Professor(a) do Magistério Superior**, em 04/10/2022, às 11:01, conforme horário oficial de Brasília, com fundamento no art. 6°, § 1°, do Decreto n° 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Murillo Guimarães Carneiro**, **Professor(a) do Magistério Superior**, em 04/10/2022, às 15:17, conforme horário oficial de Brasília, com fundamento no art. 6°, § 1°, do Decreto n° 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3957274** e o código CRC **4064A22A**.

Dedico este trabalho a Deus, aos meus familiares e amigos.

Agradecimentos

Em primeiro lugar, agradeço a Deus por me iluminar e dar forças durante todo o processo de mestrado. Agradeço a Ele por ter colocado pessoas certas nos momentos certos, por me proporcionar oportunidades, saúde e perseverança.

Também gostaria de deixar aqui meus sinceros agradecimentos ao Prof. Dr. Murillo Guimarães Carneiro e ao Prof. Dr. Luiz Gustavo Almeida Martins, por suas orientações incríveis durante todo o processo. Obrigado pelos conhecimentos, pela confiança em mim depositada e compreensões nos momentos os quais precisei.

Gostaria de agradecer aos demais membros do projeto Sistemas Nanoestruturados de Liberação Oral Sustentada para Profilaxia contra o COVID-19 e redução da transmissão do SARS-COV-2, Prof. Dr. Robinson Sabino-Silva, Prof. Dr. Bruno S. Andrade e Lucas S. Palmeira, pelas instruções e procedimentos realizados para a validação experimental dessa pesquisa. Ao Instituto Federal de Rondônia (IFRO) pelo período de afastamento que me foi concedido no qual pude me dedicar totalmente ao mestrado.

Agradeço também a minha esposa, Suelene, por conselhos, companheirismo e compreensão. Aos meus pais, Adão e Sirlene, por todo apoio que me foi dado durante toda a minha vida. Ao meu tio Edir de Abreu pelo apoio no início da graduação. Ao meu amigo Valber Lemes Zacarkim pelas palavras de apoio e incentivo no início desse processo. Ao meu amigo e colega de trabalho José Lucas Brandão e aos colegas de mestrado Lucas Vieira Murilo, Ederson Schmeing pelas trocas de conhecimentos durante o mestrado.

Por fim, agradeço à Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento do Ensino Superior (CAPES) e o Instituto Nacional de Ciência e Tecnologia em Teranótica e Nanobiotecnologia (INCT-Teranano) pelo apoio financeiro.

"A persistência é o caminho do êxito."
(Charles Chaplin)

Resumo

O design de peptídeos capazes de inibir a infecção viral tem sido considerado uma das estratégias potenciais para reduzir a transmissão do SARS-CoV-2. No entanto, a questão crítica para o design de peptídeos é o grande espaço de busca, o que torna inviável avaliar todas as possibilidades. Além disso, a maioria das análises relacionadas adota docking molecular *in silico* para selecionar potenciais peptídeos, que é uma técnica demorada e altamente dependente da estrutura molecular dos peptídeos já conhecidos e da proteína alvo. Com o objetivo de auxiliar na avaliação, descoberta e seleção de peptídeos para cálculo de docking, desenvolvemos o SAGAPEP, um framework de Algoritmo Genético Assistido por *Surrogate* capaz de encontrar peptídeos com potencial para bloquear a proteína Spike do SARS-CoV-2. O modelo *surrogate* é usado para avaliação rápida e de alta fidelidade da energia de interação entre um peptídeo e a proteína Spike, enquanto o algoritmo genético busca descobrir e selecionar peptídeos de alto potencial inspirados em princípios de genética e seleção natural. Os experimentos foram conduzidos usando um conjunto de dados composto por vários peptídeos potenciais obtidos por meio de docking molecular por especialistas em bioinformática. Como principais resultados, o SAGAPEP obteve baixas previsões de erro de seu componente *surrogate* treinado sobre esse conjunto de dados e foi capaz de descobrir e selecionar peptídeos com melhor energia de ligação do que todos listados no conjunto de dados. Além disso, os resultados notáveis do SAGAPEP sugerem que ele também pode ter o potencial de fornecer resultados promissores para outros problemas de design de peptídeos.

Palavras-chave: SARS-CoV-2, Design de Peptídeo, Aprendizado de Máquina, Modelo *Surrogate*, Algoritmo Genético, COVID-19.

Abstract

The design of peptides capable of inhibiting the SARS-CoV-2 viral infection has been considered one of the potential strategies to reduce the transmission of SARS-CoV-2. However, a critical issue in peptide design is the large search space, which makes it impracticable to evaluate all possibilities. Furthermore, most related works adopt in silico molecular docking to select potential peptides, which is a time-consuming technique and highly dependent on the molecular structure of already known peptides and the target protein. Aiming to assist the evaluation, discovery and selection of peptides for docking calculation, we developed SAGAPEP, a Surrogate-Assisted Genetic Algorithm framework capable of finding peptides with potential to block the SARS-CoV-2 Spike protein. The surrogate model is used for fast and high-fidelity evaluation of the interaction energy between a peptide and the Spike protein, while the genetic algorithm seeks to discover and select high-potential peptides inspired by principles of genetics and natural selection. Experiments were conducted using a data set composed of several potential peptides obtained through molecular docking by bio-informatics specialists. As main results, SAGAPEP achieved low error predictions from its surrogate component trained over that data set, and was able to discover and select peptides with higher binding energy than all listed in the data set. Moreover, the noteworthy results of SAGAPEP suggest it may also have the potential to provide promising results for other peptide design problems.

Keywords: SARS-CoV-2, Peptide Design, Machine Learning, Surrogate Models, Genetic Algorithms, COVID-19.

Lista de ilustrações

Lista de tabelas

Lista de siglas

AAC Composição dos Aminoácidos

AAL Ligação dos Aminoácidos

AG Algoritmo Genético

AM Aprendizado de Máquina

AS Aprendizado Supervisionado

AVPs Peptídeos Antivirais

BR Regressão Bayesiana

CKSAAGP Composição de Pares de Grupos de Aminoácidos com Espaçamento k

CKSNAP Composição de Pares de Aminoácidos com Espaçamento k

KNN K-Vizinhos Mais Próximo

MLP Perceptron Multicamadas

PSO Otimização por Enxame de Partículas

RF Florestas Aleatórias

RMSE Raiz Quadrada do Erro-Médio

S Glicoproteínas Spike

SRM Minimização do Risco Estrutural

SVM Máquina de Vetor de Suporte

SVR Regressão por Vetores de Suporte

Sumário

Capítulo 1

Introdução

Em muitos problemas do mundo real, a análise de uma solução computacional requer um grande tempo de espera ou é difícil de mensurar com precisão. Nesse cenário, a aplicação de modelos *surrogates* torna-se atrativa. *Surrogates* são modelos treinados para mensurar soluções computacionais a partir de predições aproximadas do seu valor real, com tempo de espera consideravelmente menor (MOLNAR, 2018). Várias técnicas do aprendizado de máquina (AM) são comumente aplicadas como modelos *surrogates*. O AM é um ramo da inteligência artificial que visa o desenvolvimento de modelos computacionais capazes de "aprender" padrões ou comportamentos através de observações na forma de coleção de dados (DUDA; HART; STORK, 2001; BISHOP, 2006; CARNEIRO, 2017). Os principais paradigmas de AM são: Aprendizado supervisionado, Aprendizado não supervisionado e Aprendizado por reforço.

O Aprendizado supervisionado (AS) compreende a construção de um modelo de previsão usando informações extraídas de um conjunto de dados de treinamento (CUPERTINO; ZHAO; CARNEIRO, 2015). Dessa forma, o modelo de AS é treinado através de um conjunto de dados de entrada com seus respectivos valores de saídas e, posteriormente, é utilizado para predizer os valores de saída de outros conjuntos de dados. Esses valores de saída podem assumir valores contínuos ou discretos: este denomina-se problema de *classificação* e aquele problema de *regressão*.

Em muitas situações a busca de soluções ótimas pode se deparar com um espaço de busca extremamente grande, de modo que se torna inviável a sua realização através de uma busca exaustiva. Nesse contexto, surge a necessidade da implementação de algoritmos de busca e otimização. Os algoritmos meta-heurísticos bioinspirados têm se mostrado poderosos para solução de problemas de otimização, ao abrirem mão da solução exata ótima por um conjunto de soluções aproximadas que são viáveis e que resolvem o problema em um período de tempo aceitável (YANG, 2014). Algoritmos bioinspirados são técnicas que procuram compreender os padrões encontrados na natureza para aplicá-los no desenvolvimento de ferramentas computacionais (CASTRO; ZUBEN, 2005). Entre esses algoritmos encontra-se algoritmos genéticos, otimização por colônia de formigas,

otimização por enxame de partículas e colônia de abelhas artificiais.

1.1 Motivação

O SARS-CoV-2, também conhecido como novo coronavírus, é um vírus da família coronavírus causador da doença COVID-19. Esse vírus iniciou a transmissão em Wuhan na China no início de dezembro de 2019 (HU et al., 2021) e se disseminou, de forma extremamente rápida, para os outros países, sendo transmitido de pessoa a pessoa. Isso levou a Organização Mundial de Saúde a declarar a COVID-19 como pandemia mundial no dia 11 de março de 2020 (WORLD, 2022). Desde então, diversas pesquisas estão sendo realizadas para o desenvolvimento de soluções que possam ser eficazes no combate da COVID-19.

O desenvolvimento de peptídeos como terapia eficaz contra patógenos virais emergentes é uma uma abordagem promissora devido aos seus pequenos tamanhos, são fáceis de sintetizar e têm a capacidade de penetrar nas membranas celulares. Eles também têm alta atividade, especificidade e afinidade; interação medicamentosa mínima; e diversidade biológica e química (MARQUS; PIROGOVA; PIVA, 2017; MANAVALAN; BASITH; LEE, 2022). Nesse contexto, a descoberta de peptídeo que possa interagir com uma das proteína do SARS-CoV-2 se caracteriza como uma estratégia elegante e promissora no sentido da mitigação da transmissão entre hospedeiros. Peptídeo é a ligação de duas ou mais moléculas de aminoácidos unidas por meio de uma ligação amida, denominada ligação peptídica. Embora existam cerca de 300 aminoácidos conhecidos, apenas 20 deles são frequentemente encontrados como constituintes de proteínas (NELSON; COX, 2017).

Devido aos altos custos e o tempo necessário das técnicas em métodos experimentais, a modelagem computacional por docking molecular tem desempenhado um papel importante na análise de estruturas complexas peptídeo-proteína. O docking molecular, também chamado de acoplamento molecular, permite modelar a interação entre uma pequena molécula e uma proteína, avaliando-a através de uma função de pontuação de energia (MENG et al., 2011). No entanto, essa abordagem necessita que proteína alvo e as moléculas estejam em uma estrutura tridimensional (3D) (MORRIS; LIM-WILBY, 2008), além de consumir um alto custo computacional, exigindo, em muitos casos, um longo tempo de espera para concluir a análise de interação entre um peptídeo e uma proteína específica (ZHOU et al., 2018).

Dessa forma, modelos *surrogates* podem obter valores de energia de ligação entre um peptídeo e uma proteína próximos aos obtidos por acoplamento molecular realizado por servidores específicos, com um menor tempo de espera e menor custo computacional, permitindo a avaliação de uma grande quantidade de peptídeos em um curto espaço de tempo. Outro ponto que deve ser levado em consideração é a grande quantidade de peptídeos que podem ser formados pela ligação dos 20 aminoácidos frequentemente

encontrados como constituintes de proteínas (NELSON; COX, 2017). Com exatamente 10 resíduos é possível formar $\cong 20^{10}$ peptídeos, no entanto peptídeos normalmente são compostos de 2 a 50 resíduos.

A literatura científica indica que abordagens computacionais podem ser eficientes da descoberta e projeção de peptídeos contra diversos alvos. Estudos recentes utilizaram a combinação de modelos de otimização e técnicas de AM para a descoberta de peptídeos. Yoshida et al. (2018) usaram um algoritmo evolutivo e aprendizado de máquina para explorar o espaço de sequências para a descoberta de peptídeos antimicrobianos contra *Escherichia coli*. Han et al. (2020) implementaram uma Regressão por Vetores de Suporte (SVR) para prever valores contínuos de solubilidade de proteínas a um alvo e um AG para evoluir as sequências com base nos valores preditos pelo SVR. Boone et al. (2021) usaram AS e um AG para encontrar peptídeos antimicrobianos direcionados contra a bactéria *Staphylococcus epidermidis*. Barigye et al. (2021) aplicaram AG em conjunto com uma máquina de vetor de suporte (SVM) para o desenho automático de peptídeos com possível atividade inibitória do vírus da dengue. Apesar do progresso no uso de modelos de otimização e técnicas de AM para a descoberta de peptídeos, poucos trabalhos na literatura científica relatam a aplicação dessas técnicas para descoberta de peptídeos contra o SARS-CoV-2 (KABRA; SINGH, 2021; VIELHABEN; WEICKEN; STRODTHOFF, 2021).

Nesse contexto, métodos de otimização bioinspirados podem auxiliar na descoberta e na seleção de peptídeos com potenciais de inibição do SARS-CoV-2, utilizando os valores de energias predito pelos modelos *surrogates* para guiar as evoluções. Dessa forma, essa pesquisa visa investigar a aplicação de modelos *surrogates* e algoritmos de otimização bioinspirados para a descoberta e seleção de peptídeos com potenciais de inibição da infecção e disseminação do SARS-CoV-2. Como contribuições dessa abordagem destaca-se:

- ❏ Avaliação rápida da energia de ligação entre um peptídeo e o alvo proteína. Embora vários estudos apresentem soluções demoradas para tal processo (YOSHIDA et al., 2018; BOONE et al., 2021), a abordagem proposta por essa pesquisa busca fornecer uma instantânea e de alta fidelidade avaliação de energia entre um determinado peptídeo e uma proteína alvo considerando vários descritores de atributos.

- ❏ Descoberta e seleção de peptídeos de alto potencial através de um método global de busca e otimização. A descoberta de peptídeos geralmente é realizada manualmente ou através de métodos heurísticos locais (YOSHIDA et al., 2018; HAN et al., 2020), caso contrário a abordagem proposta realiza uma pesquisa global em todo o espaço peptídico que é um mecanismo direto para evitar peptídeos ótimos locais.

- ❏ Capacidade de avaliar e descobrir peptídeos a partir de parâmetros controláveis pelo usuário, como o tamanho dos peptídeos e a proteína alvo. Estudos anteriores

limitam suas pesquisas a peptídeos de tamanho único ou de poucos resíduos (HAN et al., 2020; BARIGYE et al., 2021). Por outro lado, o método proposto por essa pesquisa ferece mecanismos flexíveis para pesquisar de forma eficiente peptídeos de diferentes tamanhos, mesmo com poucos resíduos.

1.2 Objetivos e Desafios da Pesquisa

Considerando que a pandemia de COVID-19 ainda é um problema de saúde pública global, que peptídeos podem fornecer inibição da contaminação e disseminação do vírus, o tempo de espera para concluir a análise de interação entre um peptídeo e a proteína Spike do SARS-CoV-2, os custos computacionais exigidos para calcular essa interação e a quantidade de possíveis sequências que podem ser formadas pelos vinte aminoácidos, o presente trabalho tem como objetivo principal o desenvolvimento e a aplicação de métodos de otimização bioinspirados assistidos por *surrogates* para a descoberta e seleção de peptídeo com potenciais de inibição à transmissão e disseminação do vírus. Os objetivos específicos deste trabalho são:

- ❏ Criar representações computacionais que permitem a extração de atributos em sequências peptídicas;
- ❏ Treinar modelos *surrogates* capazes de avaliarem de forma rápida e eficiente o valor de interação entre um peptídeo e a proteína Spike do SARS-CoV-2;
- ❏ Desenvolver métodos de otimização bioinspirados para busca de sequências peptídicas de tamanhos variados com potencial de inibição à transmissão e disseminação do SARS-CoV-2.

1.3 Hipótese

A nossa pesquisa pode ser dividida em duas hipóteses maiores:

- ❏ Modelos *surrogates* podem predizer a energia de interação entre peptídeos e a proteína Spike com resultados próximos àqueles obtidos pelas técnicas de acoplamento molecular do estado-da-arte, porém com menor tempo de espera.
- ❏ Modelos de otimização bioinspirados assistidos por *surrogates* são estratégias rápidas e eficientes para a busca e descoberta de peptídeos com maiores potenciais de inibição do SARS-CoV-2 em relação às principais abordagens da literatura.

1.4 Contribuições

Dentre as contribuições geradas pelo presente trabalho, destacam-se:

- ❑ Um algoritmo de otimização bioinspirado capaz de encontrar sequências com potencial de inibição frente ao SARS-CoV-2;
- ❑ Modelos *surrogates* capazes de predizer o valor de interação entre um peptídeo e a proteína Spike do vírus, possibilitando uma grande quantidade de biomoléculas em um pequeno espaço de tempo.
- ❑ Avaliação de diferentes métodos de representação de peptídeos e de diferentes métodos de aprendizado de máquina como modelos *surrogates* para a predição do valor de energia entre um peptídeo e a proteína Spike do SARS-CoV-2.

1.5 Organização da Dissertação

Os demais capítulos da dissertação foram organizados da seguinte forma:

- ❑ No Capítulo 2 são revisados os conceitos fundamentais para compreensão do trabalho, abordando tópicos relativos a pandemia de COVID-19, peptídeos antivirais, aprendizado de máquina supervisionado e algoritmos de otimização bioinspirados.
- ❑ No Capítulo 3 é apresentado o método proposto, denominado SAGAPEP, um algoritmo genético assistido por *surrogate* para a descoberta e seleção de peptídeo com potenciais de inibição à transmissão e disseminação do SARS-CoV-2.
- ❑ No Capítulo 4 são apresentados os resultados obtidos em experimentos nas avaliações e descobertas de peptídeos.
- ❑ O Capítulo 5 faz as considerações finais do trabalho, discutindo as principais contribuições da dissertação além de apresentar as limitações os principais tópicos a serem perseguidos nos trabalhos futuros.

Capítulo 2

Fundamentação Teórica

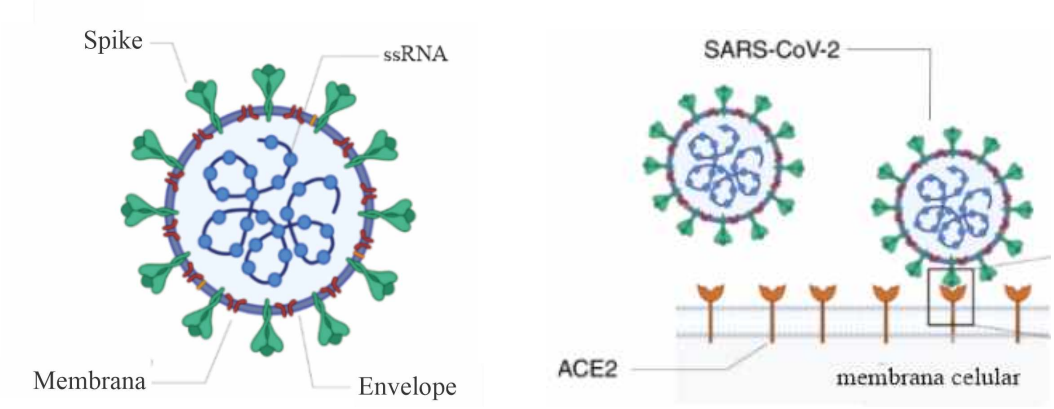
Neste capítulo serão apresentados os conceitos fundamentais e trabalhos relacionados às investigações contempladas pela dissertação. Conceitos relacionados a pandemia de COVID-19, peptídeos antivirais e acoplamento molecular são apresentados na seção 2.1. A seção 2.2 oferece uma descrição geral de tarefas e técnicas associadas ao prendizado de máquina supervisionado. A seção 2.3 introduz os principais conceitos relacionados aos algoritmos genéticos e a seção 2.4 traz os principais trabalhos arrolados da literatura.

2.1 Pandemia de COVID-19

A pandemia de COVID-19 é o resultado de uma rápida disseminação internacional de um vírus da família coronavírus, e isso, na atualidade, ainda é um problema de saúde pública global, que vem perturbando fatores sociais, econômicos, políticos, culturais, educacionais em todo o mundo. Até julho de 2022, foram atestadas mais de 574 milhões de infecções e 6,3 milhões de mortes, o que é considerado a maior pandemia do século (WORLD, 2022). O agente patológico da doença é o coronavírus 2 da síndrome respiratória aguda grave (SARS-CoV-2). A transmissão do SARS-CoV-2 pode ocorrer através de gotículas respiratórias exaladas durante a tosse, fala, espirros, respiração de um indivíduo infectado, sintomático ou assintomático. A infecção também pode ocorrer por inalação de gotículas orais ou tocando superfícies contaminadas e, em seguida, esfregar o nariz, a boca ou os olhos (CORREIA et al., 2020). Os sintomas normalmente iniciam entre 2 a 14 dias, período de incubação do vírus, e frequentemente são febre, tosse, dor de garganta, dor de cabeça, fadiga e falta de ar (OLIVEIRA et al., 2020). As complicações da contaminação variam de casos assintomáticos até a morte.

O SARS-CoV-2 é um vírus envelopado, positivo de RNA de fita simples (ssRNA [+]), protegido por uma concha de proteína nucleocapsídica e englobado por uma bicamada lipídica, conferindo-lhe uma forma esférica pleomórfica. As proteínas do envelope e membrana e as glicoproteínas Spike (S) estão ancoradas em sua superfície (OLIVEIRA et al., 2020). A proteína S é constituída pela subunidade S1 (contém o domínio de ligação ao

receptor, RBD, que é responsável por interagir com os receptores da célula hospedeira) e a subunidade S2 (dispara a fusão das membranas virais e do endossoma permitindo a entrada viral nas células hospedeiras). A ligação inicial do SARS-CoV-2 à célula hospedeira ocorre por meio de interação da proteína S com o receptor enzima conversora de angiotensina 2 (ACE2) do hospedeiro. A interação RBD/ACE2 desencadeia a endocitose de partículas virais e formação de endossomos (DING; LIANG, 2020). A ACE2 é uma proteína transmembrana expressa na superfície de diversas células do corpo, como o epitélio do sistema respiratório.



(a) Esquema de uma partícula SARS-CoV-2.

(b) Ligação às células hospedeiras.

Figura 1 – Estrutura do vírus e seu processo de ligação à célula hospedeira. (a) apresenta as principais proteínas que compões o SARS-CoV-2. (b) ilustra como o SARS-CoV-2 entra na célula através da proteína S na sua superfície por meio da ligação com seu receptor ACE2.

Fonte: Adaptado de Yang et al. (2020)

2.1.1 Peptídeos Antivirais

Mesmo com os constantes avanços nas produções de vacinas e medicamentos, os vírus continuam sendo uma das principais causas de doenças humanas. Nos últimos anos, ocorreram diversos surtos de infecções virais, como SARS, Ebola, MERS, zika vírus (ZIKV), SARS-CoV-2. Uma das alternativas mais utilizada para o controle viral é o tratamento com drogas antivirais (BOAS et al., 2019). Recentes evidências destacam a função dos compostos proteicos antivirais como barreira defensiva, os quais vêm demonstrado que alguns peptídeos antimicrobianos também podem apresentar atividade contra uma ampla gama de vírus, sendo assim, chamados de peptídeos antivirais (AVPs) (BOAS et al., 2019).

Peptídeo é a ligação de duas ou mais moléculas de aminoácidos unidas por meio de uma ligação amida, denominada ligação peptídica (NELSON; COX, 2017). Essas moléculas podem ser originadas de fontes naturais ou de fontes artificiais, quando são utilizadas

ferramentas de bioinformática, e são chamados de AVPs assinados ou artificiais. Esses AVPs podem ser derivados de estudos de "iscas" onde um peptídeo artificial é testado por interação contra um alvo específico, tal como uma glicoproteína de superfície ou uma importante enzima viral, ou obtidos *in silico* usando software específico projetado para a previsão de peptídeos. Em ambos os casos, vários aspectos podem ser levados em consideração, como topologia, composição de aminoácidos, carga e muitas outras características químicas e estruturais que podem influenciar a atividade antiviral de um peptídeo (BOAS et al., 2019).

2.1.2 Acoplamento Molecular (docking)

Acoplamento molecular se tornou uma importante abordagem para a descoberta de medicamentos. Essa abordagem pode ser usada para modelar a interação entre uma pequena molécula e uma proteína, seu objetivo é predizer a estrutura do complexo ligante-receptor usando métodos computacionais. O processo de ancoragem envolve duas etapas básicas: previsão de conformação do ligante, bem como sua posição e orientação dentro dos sítios de ligação (geralmente chamados de *pose*) e avaliação da afinidade de ligação, através de uma função de pontuação de energia *(scoring)* (MENG et al., 2011).

O *scoring* avalia a energia de ligação de interação entre o ligante e o receptor-alvo, classificando os melhores modos de ligação. Além disso, tem como propósito diferenciar poses corretas de poses incorretas, ou fichários de compostos inativos em um tempo de computação razoável (KITCHEN et al., 2004). As funções de pontuação podem ser divididas em três tipos ou classes: baseada em campo de força, empírica e funções de pontuação baseadas em conhecimento (KITCHEN et al., 2004). O processo de ligação entre um ligante e seu alvo receptor não é simples, fatores entálpicos e entrópicos influenciam na interação entre eles (ALONSO; BLIZNYUK; GREADY, 2006). Os programas de docking mais conhecidos são HPEPDOCK, AutoDock 4.0, GOLD, flexX, DOCK, entre outros.

O HPEPDOCK é um software que utiliza uma função de pontuação baseada em conhecimento iterativa para interações proteína-proteína, descrita em Huang e Zou (2008), para isso utiliza-se das estruturas 3D da proteína e do peptídeo. O software realiza o docking global e o docking local. De acordo com Zhou et al. (2018) o servidor HPEPDOCK consome uma média de 29,8 minutos para um trabalho de encaixe de peptídeo global e 14,2 minutos para um trabalho de encaixe de peptídeo local.

2.2 Aprendizado de Máquina Supervisionado

O aprendizado de máquina é um conjunto de regras utilizadas para ensinar computadores a "aprenderem" de forma automática padrões e comportamentos a partir de dados de treinamento (MOLNAR, 2018). O aprendizado supervisionado (AS) é um tipo de

aprendizagem que utiliza exemplos de pares de entrada e saída. A partir desse conjunto de dados de treinamento, deseja-se aprender uma função que faz o mapeamento da entrada para a saída. Como por exemplo, dado um conjunto de treinamento de N pares de exemplos de entrada e saída (x_1,y_1), (x_2,y_2),...(x_n,y_n), onde cada y_j foi gerado por uma função desconhecida $y = f(x)$, o AS projeta uma função h que se aproxime da função verdadeira f (RUSSELL; NORVIG, 2013). A precisão da função h é determinada durante a fase de treinamento, também conhecido como fase de aprendizagem. Uma vez que um modelo é treinado, ele pode ser utilizado para prever valores de saída de dados de entrada diferentes dos dados utilizados na fase de treinamento, a habilidade de prever corretamente dos novos valores de saída é conhecida como generalização (RUSSELL; NORVIG, 2013).

Se o resultado da previsão de um AS for uma categoria, então a tarefa é chamada de classificação, como por exemplo, a predição do conceito de um aluno em uma disciplina em uma das categorias A, B, C, D e E. No entanto, se a predição de um AS for um valor numérico específico, então a tarefa é denominada de regressão, como por exemplo, a predição do valor da nota de um aluno em uma disciplina. Algoritmos de aprendizado de máquina podem aprender por alterações de parâmetros (como pesos lineares) ou estruturas de aprendizagem (como árvores).

2.2.1 Floresta Aleatória

Floresta aleatória (*RF*, do Inglês *Random Forest*) é um método de aprendizado de máquina que é amplamente utilizado para solução de problemas de regressão e de classificação. Esse método utiliza a criação de florestas, onde cada floresta é composta por um conjunto de árvores de decisão que juntas realizam a predição do resultado (BREIMAN, 2001). Em regra, cada árvore tem igual colaboração na predição. Essa técnica de combinação de resultados de múltiplos modelos de aprendizagem de máquina (no caso da *RF* múltiplas árvores de decisão) para produzir um melhor modelo preditivo é conhecida como *Ensemble methods* (DIETTERICH, 2000).

O treinamento de cada árvore na floresta do *RF* utiliza−se de dois processos aleatórios. O primeiro é a criação de um conjunto de dados *bootstrap*, que consiste de uma técnica de amostragem de dados, para o crescimento de cada árvore na floresta. Esse conjunto de dados é composto por T amostras da base de dados de treinamento, permitindo repetição de uma mesma amostra. O segundo processo aleatório é a seleção de um subconjunto aleatórios de atributos, de modo que toda vez que for adicionar um nó na árvore, seleciona aleatoriamente um subconjuntos de A atributos para serem candidatos na escolha do atributo daquele nó, a escolha entre esses atributos é realizada pelo cálculo de impureza, esse processo não permite repetição de atributo já usados na árvore. Esses dois processos são repetidos até atingir a quantidade de árvores desejadas na floresta.

A consolidação de um resultado de uma *RF* depende do tipo problema abordado, se o problema for de regressão, a média entre os resultados das árvores é computado e dado como saída da floresta. Em caso de classificação, o resultado final é obtido através de votação dos resultados de cada árvore da floresta (HASTIE; TIBSHIRANI; FRIEDMAN, 2014).

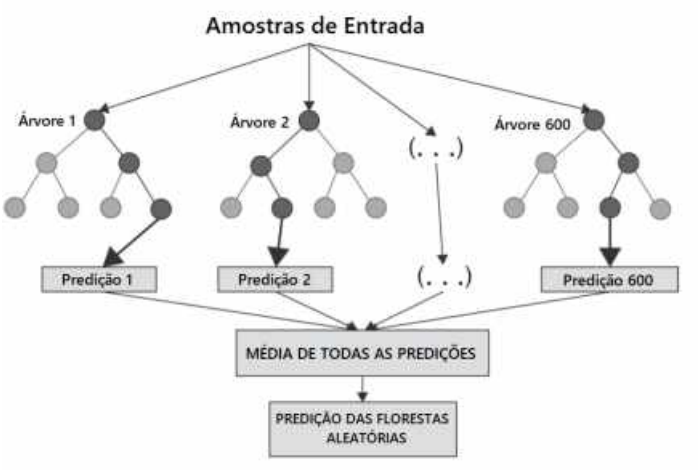


Figura 2 – Exemplo do esquema da floresta aleatória.

Fonte: Bertoni (2021)

2.2.2 Regressão Linear Bayesiana

Modelos lineares são modelos que compartilham a propriedade de serem funções lineares nos ajustes dos parâmetros. A forma mais simples de um modelo de regressão linear além de utilizar funções lineares para ajustar os parâmetros, também assumem que existe uma combinação linear dos recursos, isso é geralmente conhecido simplesmente como *regressão linear* (BISHOP, 2006), esse modelo pode ser expressado através da Equação 1.

$$y(x, w) = w_0 + w_1 x_1 + w_2 x_2 ... w_n x_n \ , \tag{1}$$

onde x = (x_1, x_1,...,x_1) são o conjunto de dados de treinamento e o vetor $w = w_1 ... w_n$ são os parâmetro de coeficientes e w_0 é um parâmetro de interceptação.

No entanto, um conjunto de modelos, conhecidos como *modelos lineares*, utilizam a linearidade apenas nas funções de ajustes de parâmetros e nos recursos de entrada utilizam a combinação de funções fixas não lineares, Equação 2.

$$y(x, w) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x) \ , \tag{2}$$

onde ϕ são conhecido como funções básicas e M o total de parâmetros do modelo. Existem diversas formas de escolhas das funções básicas, como sigmoide, gaussianas, polinômios.

Cada função de base representa uma frequência específica e tem extensão espacial infinita (BISHOP, 2006).

Neste contexto, técnicas de regressão bayesiana pode ser utilizadas para a regularização dos parâmetros. Modelos bayesianos para regressão linear utilizam distribuições de probabilidade em vez de estimativas pontuais, para isso é introduzido uma distribuição probabilística a priori sobre os parâmetros w do modelo, essas distribuições a priori pode ser dada por uma gaussiana de média zero, governado por um único parâmetro de precisão λ, como apresenta a equação 3 (TIPPING, 2001).

$$p(w|\lambda) = \mathcal{N}(w|0, \lambda^{-1}I) \ , \tag{3}$$

tendo λ como um parâmetro a ser informado.

Os modelos bayesianos tem como objetivo encontrar a distribuição posterior, dessa forma definida a priori, esses modelos procede calculando, a partir da regra de Bayes, a distribuição posterior que é condicionada às entradas e saídas de treinamento. Esse é um processo interativo, até que se atinja o número de interações desejadas ou se atinja uma margem de erro desejada. A obtenção de uma distribuição posterior pode ser obtida através da Equação 4.

$$p(w, \lambda|y, x) = \frac{p(y|w, x, \lambda)p(w, \lambda|x)}{p(y|x)} \ , \tag{4}$$

onde p(w,λ,| y,x) é a distribuição de probabilidade posterior dos parâmetros do modelo, p(w,λ|x) é a probabilidade anterior dos parâmetros, p(y|x) uma constante de normalização (MACKAY, 1992).

Desse modo, a dado um novo valor x', previsões são feitas para o alvo correspondente y' através da Equação 5 (TIPPING, 2001). Esses métodos não podem realizar os cálculos analiticamente de forma completa, portanto eles buscam uma aproximação efetiva.

$$p(y'|x') = \int p(y'|w, \lambda)p(w, \lambda|x) \tag{5}$$

2.2.3 Máquinas de Vetores de Suporte

Máquinas de vetores de suporte (SVM, do Inglês *Support Vector Machine*) são fundamentadas na teoria de aprendizado estatístico. Os SVMs utilizam o princípio de minimização do risco estrutural (*Structural Risk Minimization*, SRM). Este princípio indutivo é baseado no fato de que a taxa de erro de uma máquina de aprendizagem sobre dados de teste (i.e., a taxa de erro de generalização) é limitada pela soma da taxa de erro de treinamento e por um termo que depende da dimensão de *Vapnik-Chervonenkis* (HAYKIN, 2001). A aplicação de SVMs foi originalmente desenvolvida para problemas de classificação, sendo posteriormente estendida para o tratamento de problemas de regres-

são e denominada Regressão por Vetores de Suporte (*Support Vector Regression*, SVR) (SMOLA; SCHÖLKOPF, 2004).

O objetivo de uma SRV é encontrar uma função *f(x)* em que suas saídas são mais próximas possível dos valores reais, para isso utiliza-se uma margem de erro aceitável denominada de tubo insensível. As variáveis de saída que estão fora do tubo recebem uma das duas penalidade, dependendo se eles ficam acima ($\varepsilon+$) ou abaixo (ε-) do tubo (onde $\varepsilon+ > 0$, $\varepsilon- < 0$ para todas as amostras do conjunto de treinamento) (FLETCHER, 2008).

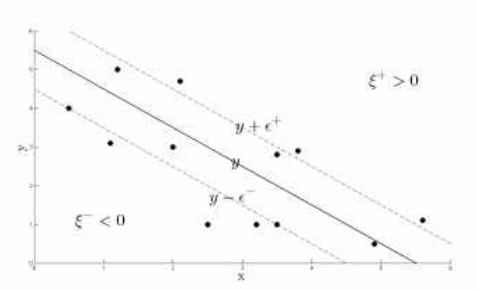


Figura 3 – Exemplo de um modelo de SVR com tubo insensível.

Fonte: Adaptado de Fletcher (2008)

No SVR, a penalidade é aumentada de acordo com a distância dos dados incorretos, ou seja, quanto mais distante do valor real maior é a penalidade. Dessa forma, o SVR tenta diminuir o número de erro, bem como suas distâncias. De acordo com (SMOLA; SCHÖLKOPF, 2004) a formulação pode ser representada como:

$$min\frac{1}{2} \parallel w \parallel^2 + C = \sum_{i=1}^{L}(\varepsilon_i^+ + \varepsilon_i^-) \ , \tag{6}$$

onde w é o valor real.

2.2.4 K-Vizinhos Mais Próximos

K-vizinhos mais próximos (*KNN*, do inglês *K-Nearest Neighbor*) fornecem funcionalidades para métodos de aprendizado não supervisionados e supervisionados. O aprendizado supervisionado baseado em vizinhos vem em dois tipos: classificação para dados com rótulos discretos e regressão para dados com rótulos contínuos. O princípio por trás dos métodos do vizinho mais próximo é encontrar um número predefinido de amostras de treinamento mais próximas em distância do novo ponto e prever o rótulo a partir delas (PEDREGOSA et al., 2011).

A distância pode ser qualquer medida métrica, como a distância euclidiana, distância manhattan, distância de minkowski. Em um problema de regressão, o *KNN* calcula a média do valores de função de seus K-vizinhos mais próximos. Para um padrão desconhecido *x'*, a regressão *KNN* pode ser calculada através da Equação 7 (KRAMER, 2013).

$$F_{KNN}(x') = \frac{1}{K} \sum_{i \in N_k(x')} y_i \ , \tag{7}$$

com o conjunto $N_K(x')$ contendo os índices dos K-vizinhos mais próximos de x'.

A regressão básica do *KNN* usa pesos uniformes, ou seja, cada ponto na vizinhança local contribui igualmente para o resultado de um ponto de consulta. No entanto, pode-se utilizar a ponderação dos pontos de tal forma que pontos próximos contribuam mais para a regressão do que pontos distante (PEDREGOSA et al., 2011).

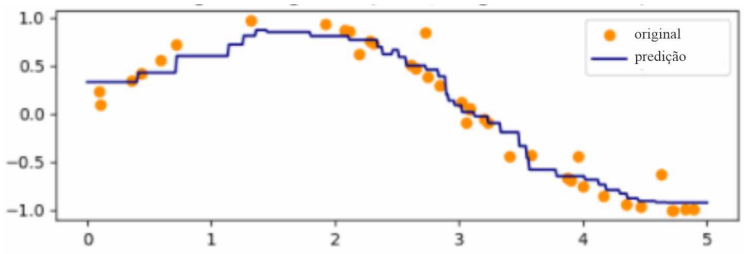


Figura 4 – Regressão *KNN* com pesos uniformes.

Fonte: Adaptado de Pedregosa et al. (2011)

2.2.5 Perceptron multicamadas

Redes neurais artificiais são interconexões massivamente paralelas de neurônio simples que funcionam como um sistema coletivo (PAL; MITRA, 1992). Um neurônio é uma unidade de processamento de informação que é fundamental para a operação de uma rede neural, cada neurônio é composto por cinco elementos básicos: um conjuntos de sinapses ou elos de conexão, cada uma caracterizada por um peso ou força própria; um somador para somar os sinais de entradas, ponderados pelas respectivas sinapses do neurônio; uma função de ativação para restringir a amplitude da saída de um neurônio; um biase para aumentar ou diminuir a entrada líquida da função de ativação, dependendo se ele é positivo ou negativo, respectivamente; e a saída, que pode ser a entrada para outro neurônio ou o resultado do problema (HAYKIN, 2001).

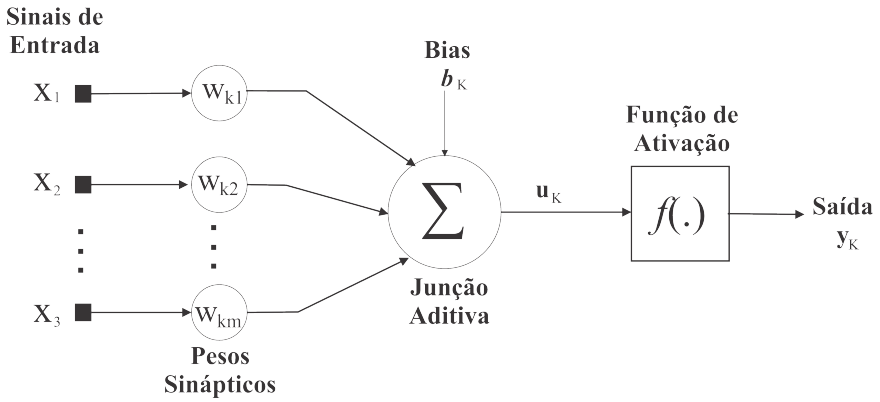


Figura 5 – Ilustração de modelo de um neurônio em redes neurais artificiais.

Uma importante classe de redes neurais é a Perceptron multicamadas (*MLP*, do inglês *Multilayer Perceptron*), esse tipo de rede consiste de um conjunto de unidades sensoriais (nós de fonte) que constituem a camada de entrada, uma ou mais camadas ocultas de nós computacionais e uma camada de saída de nós computacionais. A função dos neurônios ocultos é intervir entre a entrada externa e a saída da rede de uma maneira útil (HAYKIN, 2001). Nesse tipo de rede, os neurônios de uma mesma camada não são interconectados, no entanto todos os neurônios em uma camada estão totalmente conectados aos neurônios nas camadas adjacentes (PAL; MITRA, 1992). Esse método é utilizado para a tarefa de classificação e para a tarefa de regressão. Em se tratando de um problema de regressão o resultado é o próprio valor de saída da rede.

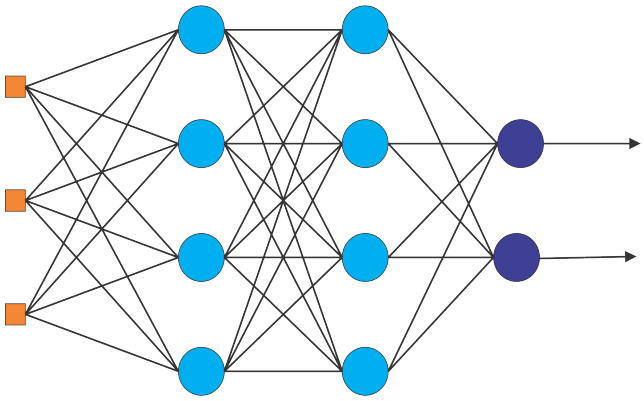


Figura 6 – *MLP* com três entradas, duas camadas ocultas e duas saídas.

O treinamento de uma *MLP* ocorre de uma forma supervisionada com o algoritmo de retropropagação do erro *(error back-propagation)*, que é baseado em uma aprendizagem por correção de erro. Basicamente, a aprendizagem por retropropagação de erro consiste de dois passos através das diferentes camadas da rede: um passo para frente, a propagação, e um passo para trás, a retropropagação. No passo para frente, um padrão de atividade (vetor de entrada) é aplicado aos nós sensoriais da rede e seu efeito se propaga através da rede, camada por camada. Finalmente, um conjunto de saídas é produzido como a resposta real da rede. Durante o passo de propagação, os pesos sinápticos da rede são todos fixos. Durante o passo para trás, por outro lado, os pesos sinápticos são todos ajustados de acordo com uma regra de correção de erro. Especificamente, a resposta real da rede é subtraída de uma resposta desejada (alvo) para produzir um sinal de erro. Este sinal de erro é então propagado para trás através da rede, contra a direção das conexões sinápticas. Os pesos sinápticos são ajustados para fazer com que a resposta real da rede se mova para mais perto da resposta desejada, em um sentido estatístico (HAYKIN, 2001).

2.3 Algoritmos Genéticos

Problemas de otimização são problemas cujo objetivo é encontrar o melhor estado de acordo com uma função objetivo (RUSSELL; NORVIG, 2013), normalmente para maxi-

mizar ou minimizar o valor da função objetivo. Por função objetivo entende-se a forma de avaliação de cada solução encontrada. Os algoritmos meta-heurísticos bioinspirados vêm se apresentado como ótimas estratégias para resolução de problemas de otimização, ao abrirem mão da solução exata ótima por um conjunto de soluções aproximadas.

Algoritmos bioinspirados são métodos meta-heurísticos que se baseiam em comportamentos observados na natureza. Esses algoritmos levam em conta dois critérios principais de qualquer algoritmo meta-heurístico: a diversificação e a intensificação. A intensificação foca na pesquisa em uma região local, explorando as regiões onde boas soluções foram encontradas, isso ocorre com a combinação das melhores soluções. A diversificação significa gerar diversas soluções para explorar o espaço de pesquisa em escala global, evitando que as soluções sejam aprisionadas em ótimos locais e, ao mesmo tempo, aumentando a diversidade das soluções (YANG, 2014).

Algoritmos Genéticos (AGs), uma das técnicas mais populares da computação bioinspirada, foram propostos e analisados primeiramente por John Holland, seus colegas e alunos na Universidade de Michigan. Os AGs são métodos de busca estocásticos baseados nos mecanismos de seleção natural e da herança genética, inspirada na teoria da evolução natural de Charles Darwin. Eles combinam estruturas de sobrevivência dos indivíduos mais aptos com estruturas aleatórias (GOLDBERG, 1989). Os AGs são frequentemente vistos como otimizadores de função, embora a gama de problemas para os quais eles foram aplicados é bastante ampla (WHITLEY, 1994).

Os AGs são iniciados a partir de um conjunto de soluções geradas aleatoriamente, denominada de população. Cada indivíduo da população é avaliado quanto a uma métrica de aptidão *(fitness)*, geralmente essa métrica é determinada através do cálculo da função objetivo, que depende das especificações do problema estudado. Um indivíduo é uma estrutura de dados que representa uma possível solução para o problema. A estrutura de um indivíduo é comumente chamada de cromossomo, eles são os genótipos que são manipulados pelo AG, e cada parâmetro do cromossomo é chamado de gene (BÄCK; FOGEL; MICHALEWICZ, 2000). A codificação binária de um indivíduo, proposta por John Holland, é uma das principais utilizadas, embora muitos também adotam outras, mais receptivas aos problemas que estão sendo enfrentados.

Após a geração e avaliação dos indivíduos da população inicial, seleciona-se alguns desses para a operação de cruzamento e mutação, normalmente essa escolha é realizada probabilisticamente, atribuindo maiores probabilidades para indivíduos com melhores aptidões (BÄCK; FOGEL; MICHALEWICZ, 2000). O cruzamento é o operador responsável por recombinar partes de dois pais para produzir um ou mais filhos, simulando o processo de troca de material genético que ocorre na natureza (BUENO, 2010). Como resultado, duas soluções são emparelhadas para a geração de novas soluções, as quais também são avaliadas quanto à métrica de aptidão. Dessa forma, os indivíduos mais aptos, dentre os existentes na população em um determinado momento, têm a tendência de passar o

material genético para as próximas gerações, assim espera-se uma melhoria da população com o passar das gerações, gerando soluções mais aptas para resolução do problema.

Após a operação de cruzamento é aplicado a operação de mutação, alterando-se o valor de um ou mais genes de um indivíduo sorteado aleatoriamente com uma determinada probabilidade, normalmente a taxa de mutação é baixa. Em alguns casos, a mutação é interpretada como gerando aleatoriamente um novo bit (WHITLEY, 1994). A função da mutação é garantir uma maior varredura do espaço de busca e evitar que o algoritmo genético convirja muito cedo para mínimos locais (LINDEN, 2012).

Depois de um novo conjunto de solução ser criado por meio dos operadores genéticos, as duas populações de pais e filhos devem ser fundidas para criar uma nova população, uma vez que a maioria dos AGs mantém uma população M de tamanho fixo, isso significa que um total de M indivíduos precisam ser selecionados a partir das populações de pais e filhos para criar uma nova população (BÄCK; FOGEL; MICHALEWICZ, 2000). Assim, o processo de seleção, cruzamento, mutação e reinserção é repetida a cada iteração do AG, até atingir o critério de parada, que normalmente é a quantidade de geração. O pseudocódigo de um AG é apresentado no Alg. 1.

Algoritmo 1 Pseudocódigo de um Algoritmo Genético

```
g ← 0
Gera uma população inicial P_g
Avalia cada indivíduo da população P_g
while condição de terminação não satisfeita do
    Crie pares de pais com algum método de seleção
    Aplique os operadores de cruzamento e mutação nos indivíduos selecionados, para
    gerar uma população de filhos Q_t
    Avalia cada indivíduo da população Q_t
    Selecione os indivíduos para a próxima geração em P_g ∪ Q_t
    g ← g+1
end
```

2.4 Trabalhos Relacionados

Métodos baseados em aprendizado de máquina têm sido usados para classificar e projetar peptídeos para lidar com diferentes tipos de doenças (KUMAR; CHAUDHARY; CHAUHAN, 2015; QURESHI; TANDON; KUMAR, 2015; AGRAWAL et al., 2021). Além disso, os trabalhos mais correlacionados também investigaram estratégias de otimização para descoberta de peptídeos (YOSHIDA et al., 2018; BOONE et al., 2021; BARIGYE et al., 2021; HAN et al., 2020). Na literatura, tais estudos envolvem a projeção e evolução de peptídeos contra os vírus *Escherichia coli*, *Staphylococcus Epidermidis* e *Dengue*.

Yoshida et al. (2018) propuseram um método baseado em otimização evolutiva e aprendizado de máquina para a descoberta de peptídeos antimicrobianos contra *Escherichia*

coli. O algoritmo genético começa gerando uma população de peptídeos candidatos (indivíduos) a partir de uma biblioteca previamente identificada. A seguir, todos os indivíduos são avaliados *in vitro* (função de aptidão). Os resultados de tais avaliações são então usados para treinar um modelo de aprendizado de máquina para prever as substituições de aminoácidos com maior potencial em um determinado peptídeo. Posteriormente, a população de indivíduos é evoluída por meio de operadores genéticos, que são guiados pelas substituições apontadas pelo modelo de aprendizado de máquina. Indivíduos com maiores avaliações *in vitro* são selecionados para a próxima geração, na qual todos os indivíduos são avaliados *in vitro* novamente e o modelo de aprendizado de máquina é retreinado com os dados gerados das gerações anteriores. Este processo é repetido até atingir um critério de parada. Diferente do nosso estudo, em que nosso modelo *surrogate* é adotado para cálculo de fitness rápido e de alta fidelidade, Yoshida et al. (2018) usa avaliação *in vitro*, enquanto modelos de aprendizado de máquina são adotados para apoiar os operadores genéticos do AG. Embora a avaliação *in vitro* seja capaz de fornecer resultados mais precisos, também é muito mais demorada e cara. Além disso, a seleção de substituições específicas de aminoácidos usando aprendizado de máquina pode limitar a exploração completa do espaço de busca levando a soluções ótimas locais.

Outro estudo baseado na avaliação *in vitro* é apresentado em Boone et al. (2021). Os autores propuseram um algoritmo genético equipado com um modelo de aprendizado de máquina para encontrar peptídeos ativos contra a bactéria *Staphylococcus epidermidis*. A população inicial do algoritmo genético é composta por peptídeos antimicrobianos conhecidos. A classificação dos dados é realizada pelo modelo de aprendizado de máquina para detectar peptídeos ativos e inativos contra a bactéria. Indivíduos inativos são descartados da população e os ativos são avaliados por ensaios *in vitro* para fornecer seus valores de fitness. Apenas 25% dos melhores indivíduos são submetidos aos operadores genéticos para gerar descendentes. Os melhores indivíduos são então selecionados para a próxima geração por meio de uma estratégia elitista. O processo é repetido até atingir o critério de parada. Apesar do uso da avaliação *in vitro*, que demanda tempo e custos consideráveis para evoluir a população, este trabalho também se diferencia do nosso pelo uso do modelo de aprendizado de máquina. Tal modelo de classificação é adotado aqui como um filtro para descartar peptídeos candidatos inativos da população. Apesar da suposta vantagem de guiar a busca por peptídeos ativos, tal estratégia também limita uma exploração completa do espaço de busca.

Na mesma direção do nosso trabalho, outros estudos também investigaram métodos de aprendizado de máquina como modelos *surrogates*. Um algoritmo genético foi aplicado para o desenho automático de peptídeos com possível atividade inibitória do vírus *Dengue* por Barigye et al. (2021). Nesse artigo, a população inicial é gerada aleatoriamente de acordo com uma das seguintes configurações: sequências de comprimento fixo com seis resíduos ou sequências variáveis com três a sete resíduos. Um modelo de classificação

de máquina de vetores de suporte é usado para fornecer a adequação de cada indivíduo. Aqui, tal valor de aptidão representa a probabilidade do indivíduo pertencer à classe ativa. A seguir, a população é submetida a operadores genéticos e a população final é totalmente substituída pelos descendentes. Este processo se repete até que um critério de parada seja alcançado e então a melhor solução candidata é retornada. Por outro lado, nosso framework pode trabalhar com peptídeos de diferentes tamanhos em vez de apenas sequências de comprimento fixo, bem como realizar avaliações de fitness por meio de um modelo de regressão em vez de levar em conta as probabilidades de um classificador. Ambos os pontos certamente contribuem para que tenhamos em nossa melhor exploração e exploração do espaço de busca respectivamente em ambas as tarefas de avaliação e descoberta de peptídeos.

Han et al. (2020) adotaram uma Regressão de Vetor de Suporte como modelo *surrogate* para prever valores contínuos de solubilidade da enzima proteica para um determinado alvo. Os autores treinaram o modelo a partir do conjunto de dados eSol (NIWA et al., Niwaa2009), que é composto por 3148 proteínas contra *Escherichia coli*. Para evoluir as sequências, foi implementado um algoritmo genético, onde o valor de fitness é previsto pelo modelo de regressão. As sequências de bases foram representadas numericamente com o descritor composição de aminoácidos (AAC) o que significa que a representação de cada peptídeo candidato do AG não é a sequência peptídica, mas um vetor AAC. Esta é uma limitação séria quando comparada ao nosso framework, pois a análise de outros descritores exigirá várias alterações em ambos os componentes: *surrogate* e algoritmo genético. Além disso, tal método tem aplicação limitada, pois os autores não mencionaram como a representação de AAC pode ser eficientemente convertida em uma de suas muitas sequências de peptídeos correspondentes.

Capítulo 3

Framework para Avaliação e Descoberta de Peptídeos

Considerando os obstáculos na busca e avaliação de peptídeos, neste capítulo apresentamos um framework denominado de SAGAPEP, um algoritmo genético assistido por *surrogate* para seleção de peptídeos, que é capaz de descobrir peptídeos e avaliar o poder de sua interação com a proteína Spike do SARSCoV-2. Para avaliação de peptídeos, utilizamos modelos *surrogates* capazes de fazer previsões rápidas e de alta fidelidade da energia de ligação entre um peptídeo e uma proteína alvo e para a descoberta de peptídeos modelamos um AG.

O restante desde capítulo está organizado da seguinte forma. A seção 3.1 traz uma visão geral sobre o framework SAGAPEP. A seção 3.2 apresenta as etapas envolvidas no aprendizado do modelo *surrogate*, sendo descrito o conjunto de dados utilizado, os métodos de extração de atributos e o treinamento dos modelos surrogates. A seção 3.3 descreve o AG utilizado para a busca de peptídeos, bem como os seus operadores genéticos de seleção, cruzamento e mutação.

3.1 Visão Geral

O desenvolvimento do SAGAPEP pode ser dividido em dois grandes módulos: avaliação peptídica e descoberta de peptídeos. A Figura 7 resume os principais passos da estrutura do SAGAPEP.

O módulo avaliação peptídica tem com finalidade prever a energia de ligação entre um peptídeo e a proteína alvo, para isso esse módulo é composto das seguintes etapas:

- **Docking molecular dos peptídeos da base de dados:** essa etapa tem como propósito a construção do conjunto de dados para o treinamento dos modelos *surrogates*. Para isso, foram selecionadas base de dados com peptídeos conhecidos e registrados na literatura e realizado o acoplamento molecular com o software HPEPDOCK das

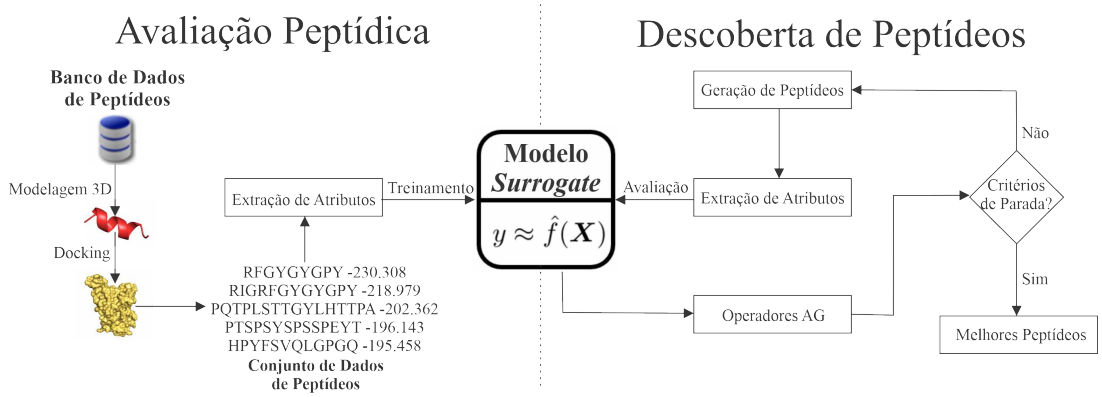


Figura 7 – Fluxograma dos módulos do framework SAGAPEP. O primeiro módulo consiste em três etapas relacionadas ao treinamento do modelo *surrogate* que será adotado para a avaliação de peptídeos. O segundo módulo é a aplicação de nosso algoritmo genético assistido por *surrogate* para a descoberta e seleção de peptídeos, um processo iterativo que é repetido até que um critério de parada seja alcançado.

moléculas com a proteína Spike para obter os valores de energia de ligação. Durante a realização do doking molecular foi definido sítio de ligação a proteína ACE2.

- **Extração de atributos:** o desempenho satisfatório dos modelos *surrogates* depende dos atributos que representam os peptídeos, essa etapa tem como finalidade representar numericamente cada molécula em um vetor ou uma matriz. Como os peptídeos dos conjunto de dados possuem tamanhos variados, isso permite a extração de atributos e padronização dos tamanhos. O SAGAPEP é composto por quatro métodos de extração de atributos descritos na literatura, os quais são descritos detalhadamente na subseção 3.2.2.

- **Treinamento dos modelos surrogates:** a etapa de treinamento da técnica proposta consiste em passar um conjunto de dados de exemplos de entrada (X_{train}) e seus respectivos valores de saída (Y_{train}) para um algoritmo de aprendizado de máquina. Aqui, utilizamos como (X_{train}) os valores de atributos extraídos pelo método de extração selecionado e como (Y_{train}) utilizamos os respectivos valores de energia de ligação *(ground-truth)* obtidos no acoplamento molecular. Os algoritmos de aprendizado de máquina utilizados são descritos na subseção 3.2.3.

O módulo descoberta de peptídeos consiste em um algoritmo genético assistido pelo *surrogate* (treinado no módulo anterior) para descobrir e selecionar peptídeos com potenciais de inibição do vírus. O SAGAPEP utiliza a estratégia de codificação de indivíduos descrita na subseção 3.3.1. A população é inicializada aleatoriamente e a aptidão de cada indivíduo é o valor predito pelo modelo *surrogate*. Os operadores de seleção, cruzamento, mutação e reinserção são descritos em seções posteriores.

As principais contribuições desta pesquisa são uma avaliação rápida da energia de ligação entre um peptídeo e a proteína alvo, a capacidade de avaliar e descobrir peptídeos de alto potencial através de um método de busca e otimização global e a partir de parâmetros controláveis pelo usuário, como o tamanho dos peptídeos.

3.2 Avaliação de Peptídeo

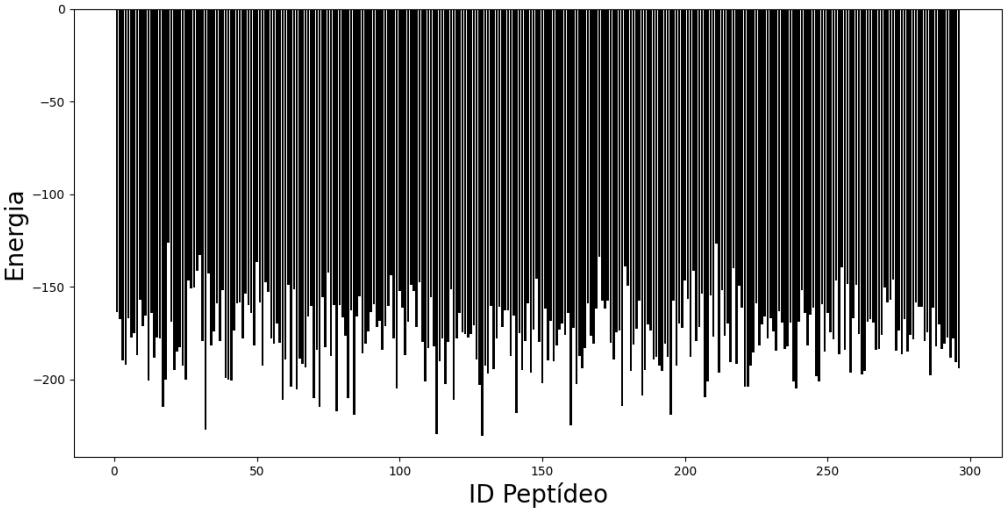
O componente de avaliação de peptídeos desempenha um papel fundamental no framework SAGAPEP, fornecendo avaliação rápida e de alta fidelidade dos peptídeos. A seguir, detalhamos as principais etapas envolvidas no aprendizado de nosso modelo *surrogate*: conjunto de dados, métodos de extração de recursos e treinamento dos modelos de aprendizado de máquina.

3.2.1 Conjunto de Dados

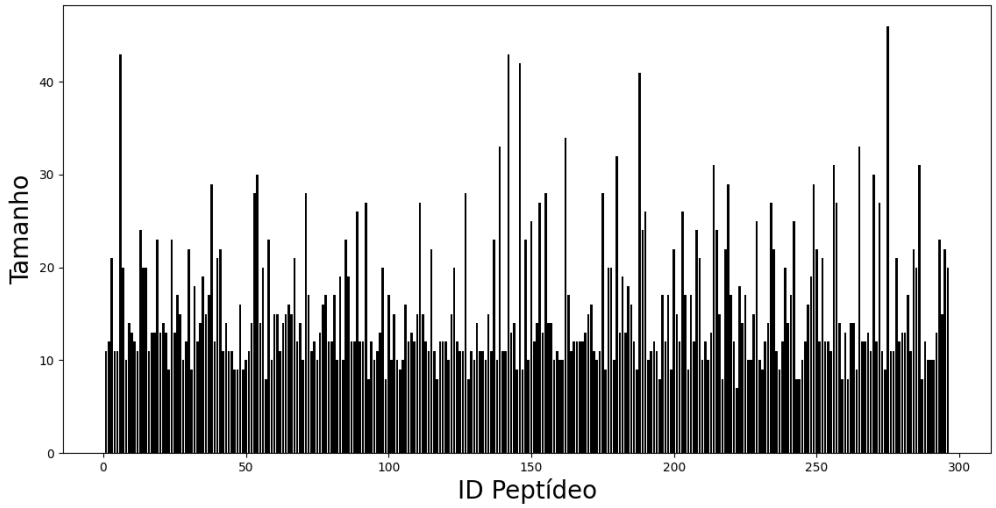
O conjunto de dados é composto por 296 peptídeos retirados de Caseiro et al. (2009) e seus respectivos valores de energia de interação com a proteína Spike do SARS-CoV-2. Os peptídeos são representados linearmente (vetor de aminoácidos). Por exemplo, no peptídeo "RFGYGYGPY" cada caractere representa um aminoácido específico, sendo R Arginina, F Fenilalanina, G Glicina, Y Tirosina e P Prolina. Primeiramente foram gerados os modelos tridimensionais dos peptídeos usando o software PEP-FOLD3 (NÉRON et al., 2013) e, posteriormente, foram realizadas as avaliações usando o software de acoplamento molecular HPEPDOCK (ZHOU et al., 2018). O resultado do acoplamento molecular é a energia de ligação dos peptídeos com a proteína Spike, por exemplo, a energia de afinidade entre o "RFGYGYGPY" peptídeo e a proteína Spike é -230,308 Kcal/Mol. Quanto menor esse valor, melhor é a interação do peptídeo contra o alvo. Os valores das energias de ligaçãos dos peptídeos no conjunto de dados variam de -126,060 Kcal/Mol a -230,308 Kcal/Mol e o tamanho dos peptídeos varia de 7 a 46 aminoácidos, tendo uma média em torno -175.975 Kcal/Mol e de 16 resíduos. A Figura 8 apresenta o histograma das energias de ligação dos peptídeos do conjunto de dados e a quantidade de resíduos.

3.2.2 Métodos de Extração de Atributos

Visando projetar um conjunto apropriado de características das sequências peptídicas disponíveis em nosso conjunto de dados, quatro métodos de extraçaõ de atributos foram implementados: composição dos aminoácidos (AAC), ligação dos aminoácidos(AAL), composição de pares de aminoácidos com espaçamento k (CKSNAP) e composição de pares de grupos de aminoácidos com espaçamento k (CKSAAGP). Além desses quatro métodos, também foram avaliados métodos híbridos baseados na junção de atributos de pares de descritores. Tais descritores são formalmente apresentados a seguir:



(a) Energias de ligação com a proteína Spike de cada peptídeo



(b) Quantidades de resíduos de cada peptídeo.

Figura 8 – Histograma dos peptídeos da base de dados usada no treinamento dos modelos *surrogates*. (a) apresenta os valores de energias de ligação dos peptídeos com a proteína Spike e (b) a quantidade de resíduos presente em cada peptídeos.

- **Composição dos Aminoácidos (AAC):** A tradicional composição de aminoácidos(AAC) em proteínas ou peptídeos contém 20 componentes, cada um refletindo a frequência de ocorrência de um dos 20 aminoácidos nativos em uma proteína (CHEN; SHEN; ZOU, 2012; KUMAR; CHAUDHARY; CHAUHAN, 2015). O AAC pode ser calculado através da equação a seguir:

$$AAC(a) = \frac{1}{N}\sum_{i=1}^{N}\delta(a, P_i) \ , \tag{8}$$

$$\delta(i,j) = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

onde a representa o aminoácido natural em análise, P_i o i-ésimo aminoácido da sequência peptídica, N é o tamanho da sequência peptídica e δ é a função Kronecker definida por (9) que retorna 1 se ambas as variáveis i e j são iguais e 0 caso contrário. Por exemplo, o AAC(G) do peptídeo RFGYGYGPY é igual a 0,33. O descritor AAC foi aplicado com sucesso para classificação de receptores nucleares (BHASIN; RAGHAVA, 2004), previsão de peptídeos anticancerígenos (WEI et al., 2018) e predição de peptídeos anti-hipertensivos (KUMAR; CHAUDHARY; CHAUHAN, 2015).

- **Ligação dos Aminoácidos (AAL):** Esse descritor reflete o modo de acoplamento entre cada resíduo com o seu resíduo adjacente, representando o modo de correlação de primeira camada do método de Chou (2000). Aqui cada peptídeo é representado por uma matriz 20×20, onde cada célula é o acoplamento de um resíduo com outro ou com si mesmo. A Figura 9 mostra uma representação ilustrativa deste descritor, que é calculado da seguinte forma:

$$AAL(a1,a2) = \sum_{i=1}^{N-1} \delta((a1,a2),(P_i,P_{i+1})) \ , \tag{10}$$

onde $a1$ e $a2$ representam os resíduos adjacentes sob análise, (P_i e P_{i+1}) refere-se a pares da sequência peptídica, N é o tamanho da sequência, e δ é a função de Kronecker definida por (9). Por exemplo, o AAL(G,Y) da sequência peptídica RFGYGYGPY é igual a 2. O método de Chou foi aplicado com sucesso a atributos celulares de proteínas predição (CHOU, 2001a) e predição de classes de subfamília de enzimas (CHOU, 2001b).

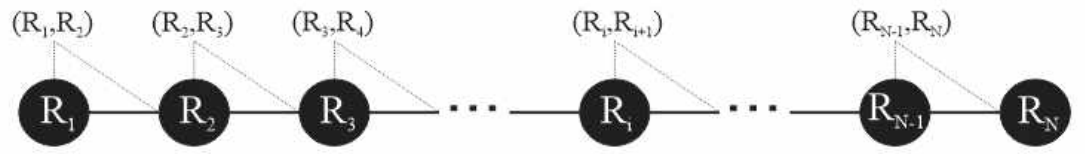


Figura 9 – Extração de recursos através da ligação de aminoácidos. Cada aminoácido se liga ao seu vizinho a direita, exceto o último resíduo (C-terminal) que não tem vizinho a direita.

Fonte: Adaptado de Silva et al. (2022)

- **Composição de Pares de Aminoácidos com Espaçamento *k* (CKSNAP):** O método CKSNAP calcula a frequência de pares de resíduos separados por qualquer ácido nucleico k tal que $k \in \{1, 2, \ldots, 5\}$ (CHEN et al., 2020). A representação deste

descritor é uma matriz 20×20, onde cada célula representa a frequência normalizada de cada par de aminoácidos no peptídeo, como segue:

$$CKSNAP(a1, a2, k) = \frac{1}{N-1} \sum_{i=1}^{N-1} \max_{\substack{1 \leq j \leq k \\ i+j \leq N}} (\delta((a1, a2), (P_i, P_{i+j}))) \ , \qquad (11)$$

em que $a1$ e $a2$ representam os resíduos adjacentes sob análise, P_i e P_{i+1} referem-se a pares da sequência peptídica, N é o tamanho da sequência peptídica e δ é a função Kronecker definida por (9). Por exemplo, o CKSNAP(G, Y, 2) do peptídeo RFGYGYGPY é igual a 0,375. O descritor CKSNAP foi aplicado com sucesso para predição de sítios de propionilação de lisina (JU; HE, 2017), identificação de proteínas antioxidantes (ZHAI et al., 2020) e predição de sítios de citrulinação (JU; WANG, 2018).

- **Composição de pares de grupos de aminoácidos com espaçamento k (CKSAAGP):** O extrator de atributos CKSAAGP calcula a frequência de ligação entre os grupos de aminoácidos (CHEN et al., 2020). Para isso, os 20 aminoácidos são divididos em 5 grupos de acordo com suas propriedades físico-químicas: GAVLMI, FYW, KRH, DE, STCPNQ (CAO RUIFEN AMD WANG; BIN; ZHENG, 2021). Levando em consideração k=1, há 25 pares de grupos, ou seja, g1g1, g1g2, g1g3,...,g5g5. O CKSAAGP pode ser calculado através da equação a seguir:

$$CKSAAGP(g1, g2, k) = \frac{1}{N-1} \sum_{i=1}^{N-1} \max_{\substack{1 \leq j \leq k \\ i+j \leq N}} (\delta((g1, g2), (T_i, T_{i+j}))) \ , \qquad (12)$$

em que $g1$ e $g2$ representam os grupos de resíduos adjacentes sob análise, T_i e T_{i+1} referem-se a pares dos grupos de peptídeos, N é o total ligações entre os grupos no peptídeo e δ é a função Kronecker definida por (9). Esse extrator foi aplicado com sucesso para identificar locais de succinilação de proteínas[1] (KAO et al., 2020) e para previsão de locais de ubiquitinação (CHEN et al., 2011).

- **Modelo Híbrido:** Nesse modelo, os atributos de dois métodos de extração de recursos são combinados para caracterizar um peptídeo. No SAGAPEP foi implementado todas as combinações dois a dois dos métodos de extração de atributos descritos anteriormente.

3.2.3 Treinamento dos Modelos *Surrogates*

Os sistemas de docagem molecular computacional demandam um alto custo, bem como um longo tempo de espera para calcular a energia de ligação entre um peptídeo e

[1] A succinilação de proteínas é uma reação bioquímica na qual um grupo succinil (-CO-CH2-CH2-CO-) é ligado ao resíduo de lisina de uma molécula de proteína.

uma proteína alvo. Por exemplo, o servidor HPEPDOCK leva um média de 14,2 minutos para um trabalho de encaixe de peptídeo local e 29,8 minutos para um trabalho global de ajuste de peptídeo (ZHOU et al., 2018). Por outro lado, a aplicação de modelos *surrogates* para a avaliação de peptídeos pode permitir uma análise instantânea e de alta precisão na avaliação de sua energia de ligação com a proteína alvo.

Os *surrogates* são modelos treinados para fazer previsões aproximadas de custos, funções e processos complexos ou de longa duração (MOLNAR, 2018). Métodos de aprendizado de máquina têm sido aplicados com sucesso como modelos *surrogates* para vários domínios, como design de vidro óptico (CASSAR; SANTOS; ZANOTTO, 2021), previsão de rendimento de bio-óleo (ULLAH et al., 2021), otimização de forma aerodinâmica (ZHANG et al., 2021), triagem de drogas contra *Mycobacterium tuberculosis* (GUPTA; BHAKTA, 2012), otimização de projeto eletromagnético (AKINSOLU et al., 2019) e assim por diante. Para prever com eficiência o valor de energia de ligação entre peptídeos e a proteína Spike de SARS-CoV- 2 foram implementados os seguintes modelos de regressão no framework: florestas aleatórias (RF), regressão bayesiana (BR), máquina de vetor de suporte (SVM), K-vizinhos mais próximos (KNN) e perceptron multicamadas (MLP).

3.3 Descoberta de Peptídeo

O Algoritmo Genético (AG) é um método de busca bem conhecido baseado nos princípios de seleção natural e genética (GOLDBERG; HOLLAND, 1988). Sua heurística baseada em população visa encontrar boas soluções para problemas de otimização. O AG composto no SAGAPEP visa descobrir peptídeos com potenciais para inibir a proteína Spike o vírus SARS-CoV-2 guiado pela avaliação de peptídeos assistida por *surrogate* apresentada anteriormente. A Figura 10 ilustra o fluxo básico do AG, que consiste em gerar um população de soluções aleatórias (sequências peptídicas) com uma evolução iterativa subsequente desses indivíduos, com base no ranking obtido a partir da avaliação assistida por *surrogates* e por meio de operadores genéticos de seleção, cruzamento e mutação, os quais são descritos com detalhes nas próximas seções.

3.3.1 Estratégia de representação e inicialização dos indivíduos

A representação de um indivíduo no AG consiste em um vetor $\mathcal{I} = \{P_1, \ldots, P_n\}$, em que P_i denota o aminoácido na i-ésima posição e n o número de aminoácidos presentes no peptídeo, tal que $2 \leq n \leq pep\_max\_size$, onde pep_max_size é um parâmetro configurável. Assim, são gerados peptídeos de diferentes tamanhos, garantindo a diversidade entre os indivíduos da população do AG e possibilitando a busca de peptídeos com poucos resíduos, o que inclusive influência na complexidade de análise *in vitro* (JOHANSSON-ÅKHE; MIRABELLO; WALLNER, 2018). Cada posição no vetor que representa um

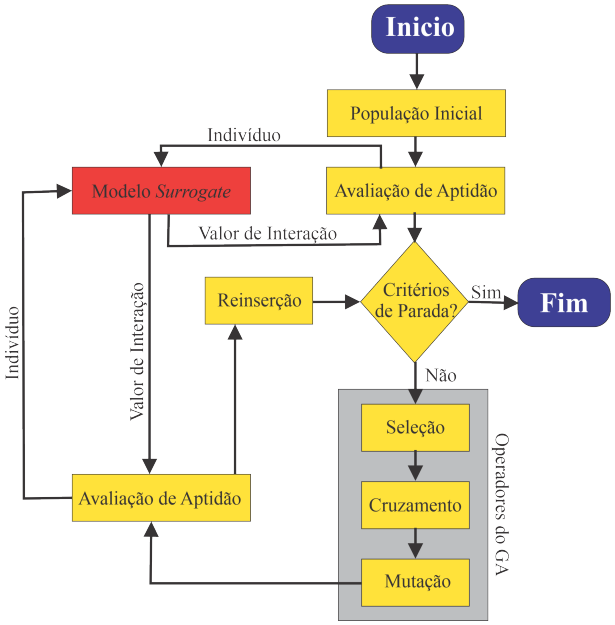


Figura 10 – Fluxo de algoritmo genético assistido por *surrogate*. O processo se inicia com uma população inicial de indivíduos, estes têm seus valores de aptidão previstos pelo modelo *surrogate*, posteriormente, o operadores genéticos de seleção, cruzamento e mutação são aplicados. Os novos indivíduos são então avaliado pelo modelo *surrogate*, então os melhores entre pais e filhos são selecionados para a próxima geração, esse processo iterativo ocorre até atingir o critério de parada

cromossomo do indivíduo armazena uma letra que reflete um dos 20 aminoácidos frequentemente encontrados como constituintes de proteínas. O aminoácido de cada posição é escolhido aleatoriamente de modo que cada um dos 20 aminoácidos tem a mesma probabilidade de ser selecionado para qualquer posição peptídica. A Tabela 1 apresenta os aminoácidos e a letra correspondente e a Figura 11 demostra um exemplo de representação de um indivíduo.

Tabela 1 – 20 aminoácidos frequentemente encontrados como constituintes de proteínas e as respectivas letras correspondentes.

Letra	Aminoácido	Letra	Aminoácido	Letra	Aminoácido	Letra	Aminoácido
A	Alanina	G	Glicina	M	Metionina	S	Serina
C	Cisteina	H	Histidina	N	Asparagina	T	Treonina
D	Aspartato	I	Isoleucina	Q	Glutamina	V	Valina
E	Glutamato	K	Lisina	P	Prolina	W	Triptofano
F	Fenilalanina	L	Leucina	R	Arginina	Y	Tirosina

Na inicialização da população são gerados uma conjunto de *pop_size* indivíduos, onde *pop_size* é um parâmetro configurável, essa população é então evoluída ao longo das gerações para encontrar peptídeos com melhores potenciais de interação com a proteína alvo.

Figura 11 – Codificação de indivíduo proposto para o AG que compõe o SAGAPEP. Cada posição armazena uma letra correspondente a um aminoácido. No exemplo, o indivíduo é composto por 9 aminoácidos.

3.3.2 Cálculo de aptidão

Para a avaliação de aptidão de cada indivíduo, um vetor de atributos é calculado diretamente da codificação individual utilizando o método de extração de atributos selecionado durante a fase de treinamento do *surrogate*. Os métodos de extração de atributos são descritos na subseção 3.2.2. Esse vetor de atributos é então inserido como entrada para o modelo *surrogate* que no mesmo instante retorna a energia de ligação prevista para essa sequência peptídica contra a proteína alvo. Esse valor de retorno é o valor de aptidão do indivíduo. Embora esta abordagem requer cálculo vetorial na avaliação de cada indivíduo, ela garante que a ordem dos aminoácidos no peptídeo será prontamente conhecida ao final do processo de otimização.

3.3.3 Seleção

Torneio simples de indivíduos de tamanho *tour* é adotado para a seleção de pares de soluções candidatas ao cruzamento. Neste tipo de seleção, são selecionados aleatoriamente um grupo de *tour* indivíduos ($tour \geq 2$) tomados da população para participar de um torneio. O vencedor desse torneio é o indivíduo com a melhor aptidão dentre os participantes. Assim, o indivíduo que vencer esse torneio é selecionado para a operação de cruzamento e fica inapto a participar de um novo torneio, enquanto os demais indivíduos que participaram da competição são descartados e poderão participar de novos torneios. Em cada operação de cruzamento é realizado dois torneios para selecionar os dois indivíduos para a reprodução.

3.3.4 Cruzamento

No cruzamento, os indivíduos filhos são gerados a partir da combinação dos cromossomos de dois pais. O SAGAPEP utiliza o cruzamento de dois pontos, nesse modelo de cruzamento dois valores são escolhidos aleatoriamente entre 0 e o tamanho do menor peptídeo selecionado, caso o segundo valor seja igual ao primeiro é realizado novo sorteio até que esse segundo valor seja diferente do primeiro, isso evita filhos idênticos aos pais. O primeiro ponto consiste no menor valor sorteado e o segundo ponto o maior valor.

Após a definição dos pontos é realizado a formação dos filhos da seguinte forma:

- **Primeiro filho:** recebe do primeiro pai os cromossomos da posição 0 até a posição do menor ponto, além dos cromossomos do maior ponto até o tamanho do primeiro

pai; do segundo pai ele recebe os cromossomos entre os dois pontos.

- **Segundo filho:** recebe do segundo pai os cromossomos da posição 0 até a posição do menor ponto, além dos cromossomos do maior ponto até o tamanho do segundo pai; do primeiro pai ele recebe os cromossomos entre os dois pontos.

O processo de cruzamento ocorre até atingir a taxa de cruzamento previamente configurada. A Figura 12 ilustra a operação do cruzamento de dois pontos utilizado no SAGAPEP com dois indivíduos de tamanhos diferentes, o primeiro pai é composto por 21 cromossomos e o segundo por 9 cromossomos. Os valores dos pontos sorteados na ilustração são 2 e 7, dessa forma são gerados dois filhos, sendo o primeiro composto por 21 cromossomos e o segundo com 9 cromossomos.

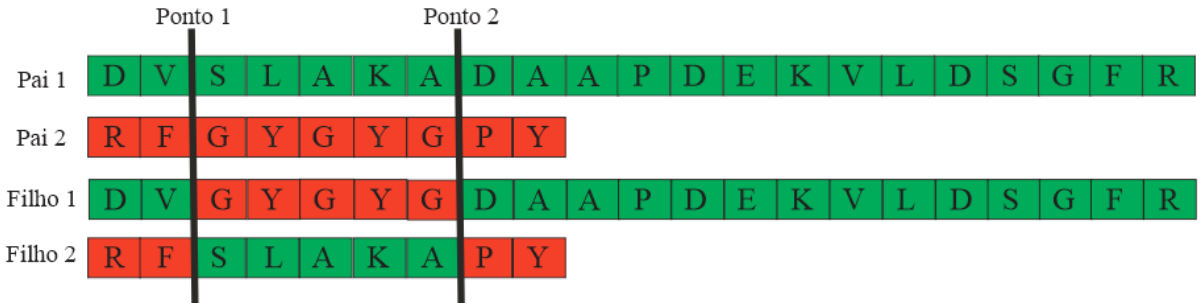


Figura 12 – Exemplo de aplicação do cruzamento de dois pontos utilizado no SAGAPEP. Através do cruzamento de dois pais gera-se dois filhos compostos por material genético de ambos os pais.

3.3.5 Mutações

Dois tipos de operadores independentes são utilizados na mutação dos indivíduos:

- **Substituição de aminoácidos:** consiste em selecionar de forma aleatória um indivíduo entre os filhos gerados e selecionar duas de suas posições aleatoriamente e substituir o valor de aminoácido presente nessas posições por dois aminoácidos sorteados, também de forma aleatória, entre os 20 aminoácidos possíveis apresentados na Tabela 1.

- **Permutação de aminoácidos:** é selecionado de forma aleatória um indivíduo entre os filhos gerados, podendo ser o mesmo da primeira mutação, ou não, e em seguida duas posições são escolhidas aleatoriamente e é realizado a permuta dos aminoácidos presente nessas posições.

Esse processo é repetido até atingir uma taxa de mutação previamente configurada. A Figura 13 ilustra os processos de mutação.

(a) Substituição de aminoácidos. (b) Permutação de aminoácidos.

Figura 13 – Estrutura de mutação proposta. (a) demostra a mutação substituição de aminoácidos onde foi sorteado as posições 4 e 8 e os aminoácidos W e C para substituir os aminoácidos L e A. (b) ilustra a mutação permutação de aminoácidos onde as posições 2 e 9 foram selecionadas para a permutação.

3.3.6 Reinserção

O operador de reinserção é responsável por definir a população da nova geração através de alguma estratégia de seleção dos indivíduos da geração atual e dos seus filhos. Após as operações genéticas de seleção, cruzamento e mutação, todos os descendentes são avaliados e a reinserção ordenada é aplicada para selecionar os indivíduos que irão compor a população da próxima geração. Na reinserção ordenada, os melhores indivíduos entre pais e filhos são mantidos próxima geração e os demais são descartados.

3.3.7 Gerações

As operações de seleção, cruzamento, mutação e reinserção ocorrem de forma iterativa, sendo realizadas até alcançar um número máximo de gerações (critério de parada).

Capítulo 4

Experimentos e Análise dos Resultados

Neste capítulo serão apresentados os experimentos realizados nesta pesquisa e os resultados obtidos. Os experimentos são divididos em duas grandes partes: avaliação peptídica e descoberta de peptídeos. Na primeira, são apresentadas as métricas de desempenho utilizadas, as configurações de hiperparâmetros e os resultados obtidos com o treinamento de vários modelos *surrogates* usando os métodos de extração de atributos descritos na subseção 3.2.2. Na segunda, é apresenta a análise de convergência do AG, bem como os peptídeos descobertos pelo SAGAPEP e seus resultados após análise por especialistas.

4.1 Avaliação Peptídica com *Surrogate*

Nesta seção será detalhado o ambiente experimental em termos de métrica utilizada para análise dos modelos *surrogates*, parâmetros configurados nos algoritmos e os resultados obtidos nos experimentos do módulo avaliação peptídica.

4.1.1 Métrica de Avaliação

A métrica utilizada para avaliação dos modelos *surrogates* é a validação cruzada repetida com 10 iterações e 10 partições. Nesta métrica, a cada iteração as sequências peptídicas do conjunto de dados são divididas aleatoriamente em dez partes. Nove delas são usadas no treinamento e a outra para teste. O processo citado é repetido dez vezes a cada interação, até que cada uma das dez partes seja usada uma vez para teste. Essa métrica produz diferentes divisões em cada iteração. A Figura 14 ilustra como é realizada a divisão a cada iteração. O desempenho do modelo é calculado com base na média do valor do raiz quadrada do erro-médio (RMSE) da seguinte forma:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{p_i - t_i}{\sigma_i}\right)^2} \tag{13}$$

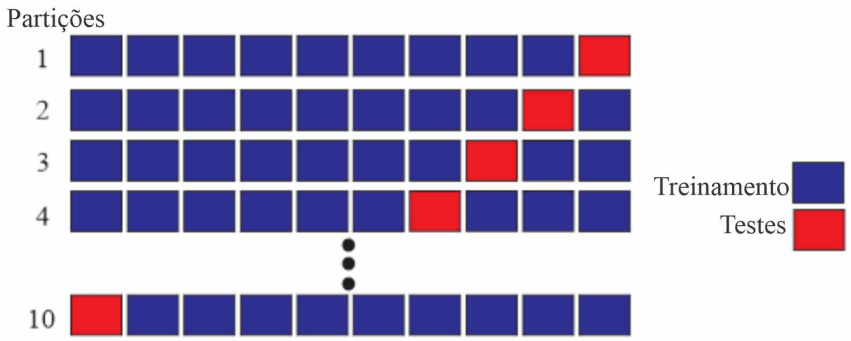


Figura 14 – Exemplo de iteração na validação cruzada com 10 partições; cada partição é trocado um conjunto de treino para teste de forma que todas as partes tenham sido usados para treinar e avaliar o modelo no processo.

4.1.2 Experimentos dos Modelos *Surrogates*

A fim de obter modelos de regressão de alta fidelidade capazes de prever com eficiência o valor da energia de ligação de um peptídeo com a proteína Spike do SARS-CoV-2, o algoritmo *Grid Search* foi utilizado para realizar a seleção de parâmetros para os métodos de aprendizado de máquina. A partir destes experimentos, chegou-se aos seguintes parâmetros de configuração: no BR, o parâmetro de normalização foi ajustado para True; no RF, o número de árvores na floresta foi configurado como igual a 200; no KNN, o número de vizinhos foi definido como igual a 10; no SVM, kernel definido como a função polinomial, com grau da função kernel polinomial igual a 3 e o parâmetro de penalidade C igual a 1; no MLP, o número de neurônios na camada oculta (100,300,500), tanh foi definido como a função de ativação e o número de épocas foi fixado em 300. No método de extração de atributos CKSNAP o valor de k foi definido como $k = 5$ e no método CKSAAGP como $k = 1$. Os resultados para outros valores de k podem ser consultados no Apêndice A.

A Tabela 2 mostra o desempenho preditivo alcançado por cada modelo *surrogate* de acordo com a técnica de regressão e método de extração de atributos considerados no treinamento do modelo. Os resultados são apresentados em termos de RMSE (média e desvio padrão) e mostram que três das cinco técnicas obtiveram melhores resultados com o método AAC e as outras duas com o método AAL. Em relação aos modelos de regressão, o melhores desempenhos foram obtidos pelos métodos BR e RF. Os modelos que obtiveram melhores avaliações são o BR/AAL e RF/AAC ambos com um RSME de -14.1. Dessa forma, a adoção de um modelo *surrogate* construído usando a configuração BR/AAL ou RF/AAC é bastante atraente, pois permitem o cálculo rápido da energia de interação das sequências peptídicas e também obter resultados de alta fidelidade, com uma taxa de erro aceitável para o módulo de descoberta de peptídeos.

A fim de verificar a eficácia dos modelos *surrogates* utilizando combinação de atributos dos extratores de atributos, também foram realizados experimentos juntando atributos de cada par de descritores, denominado atributos híbridos. Os resultados são apresentados

Tabela 2 – Desempenho dos modelos *surrogates* em termos de RMSE (média e desvio padrão) com os quatro métodos de extração de atributos.

Modelo *Surrogate*	AAC	AAL	CKSNAP	CKSAAGP
BR	15.3 ± 2.1	**14.1 ± 2.6**	14.5 ± 1.9	16.0 ± 1.9
RF	**14.1 ± 2.0**	16.7 ± 2.4	15.0 ± 1.9	14.3 ± 1.9
KNN	17.2 ± 1.9	19.1 ± 2.4	20.7 ± 2.2	17.7 ± 2.5
SVM	14.5 ± 2.0	15.6 ± 2.0	14.9 ± 2.2	15.3 ± 2.0
MLP	19.2 ± 2.2	15.9 ± 2.1	19.2 ± 2.2	19.2 ± 2.2

na Tabela 3. Nessas combinações, o melhor resultado foi obtido utilizando os extratores de atributos AAL com CKSAAGP usando o modelo de regressão BR, seguido pela combinação dos extratores de atributos AAC com AAL usando o modelo de regressão RF. Esses valores são poucos superiores aos obtidos com as configuraçoes AAL/BR e AAC/RF, dessa forma a análise dessas configurações no módulo descoberta de peptídeos se torna atraente para confirmar as suas eficácias.

Tabela 3 – Desempenho dos modelos *surrogates* em termos de RMSE (média e desvio padrão) com modelos híbridos de extração de atributos.

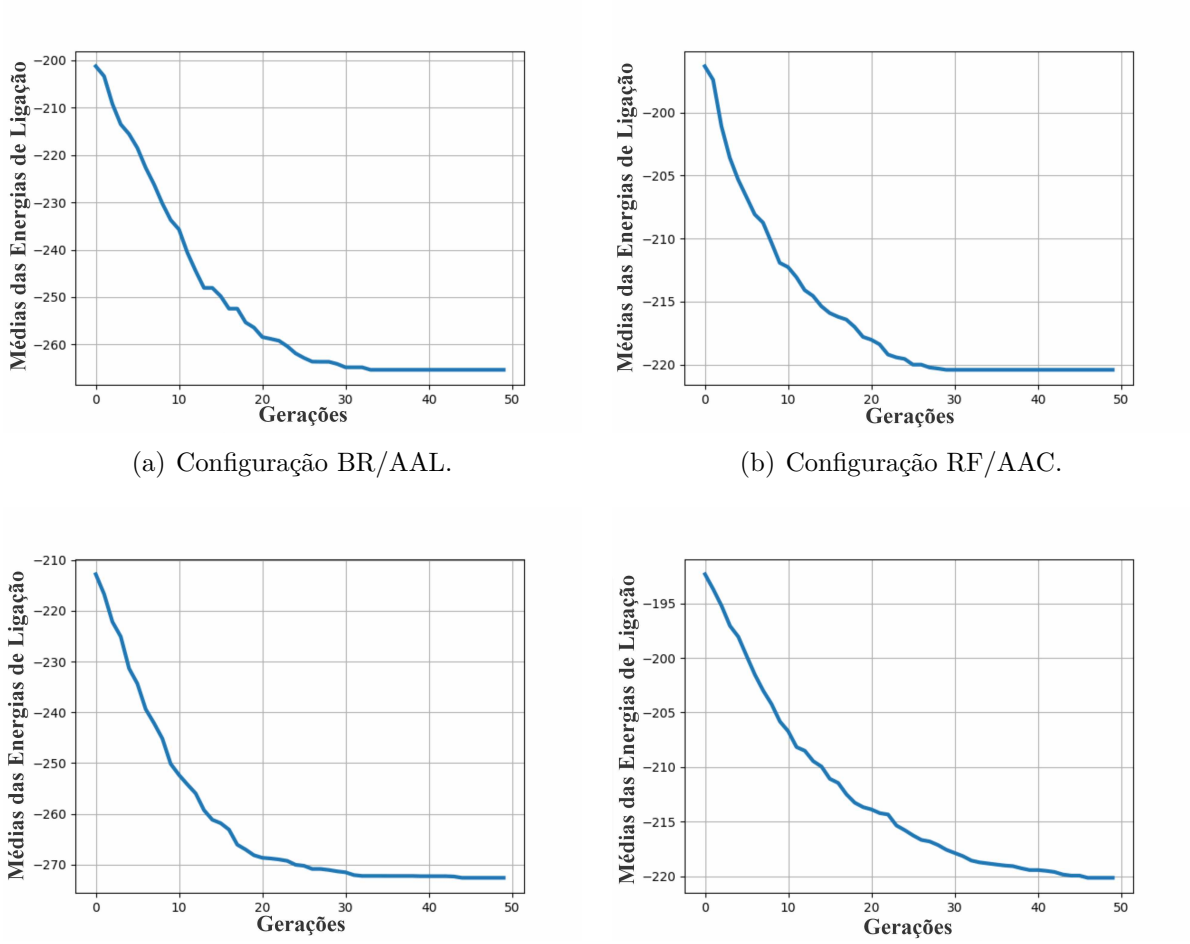
Atributos	BR	RF	KNN	SVM	MLP
AAC + AAL	13.9 ± 2.0	**13.8 ± 2.1**	19.1 ± 2.8	15.4 ± 2.1	15.9 ± 2.1
AAC + CKSNAP	14.5 ± 1.9	17.9 ± 2.1	14.3 ± 1.9	14.4 ± 2.0	19.2 ± 2.0
AAC + CKSAAGP	14.9 ± 2.2	13.5 ± 1.8	17.4 ± 2.3	13.9 ± 2.0	18.9 ± 2.4
AAL + CKSNAP	15.4 ± 2.1	15.0 ± 2.0	19.0 ± 2.5	15.6 ± 2.0	16.0 ± 2.2
AAL + CKSAAGP	**13.7 ± 1.8**	14.2 ±1.5	18.9 ± 2.5	-15.7 ± 2.1	15.4 ± 1.9
CKSNAP + CKSAAGP	14.3 ± 1.7	14.1 ± 1.8	18.1 ± 2.3	14.4 ± 1.8	19.2 ± 2.3

4.2 Descoberta de Peptídeos com SAGAPEP

Para identificar peptídeos com alto potencial de interação com a proteína Spike SARS-CoV-2, realizamos uma simulação de 10 execuções para os dois melhores modelos *surrogates* que usam apenas um extrator de atributos, sendo o BR/ALL e o RF/AAC, e os dois melhores modelos que usam atributos híbridos, sendo BR/AAL+CKSAAGP e o RF/AAC+AAL. O número máximo de resíduos nos peptídeos foi definido como $pep\_max\_size = 10$. Os parâmetros restantes do AG foram definidos da seguinte forma: tamanho da população $pop\_size = 50$, tamanho do torneio $tour\_size = 3$, taxa de cruzamento $cross\_rate = 100\%$, taxa de mutação $mut\_rate = 30\%$ e número de gerações $max\_gen = 50$.

A primeira análise teve como objetivo avaliar a convergência do nosso AG em relação aos parâmetros selecionados. A Figura 15 demonstra que SAGAPEP é capaz de evoluir para melhores peptídeos candidatos ao longo de gerações. Na figura, foi calculado a

média do melhor peptídeo por geração considerando as 10 execuções de cada configuração simuladas com nosso AG. A figura mostra um bom *trade-off* entre exploração e explotação. Até a 20ª geração houve uma melhoria de aproximadamente 60 Kcal/Mol com BR/AAL e de quase 30 Kcal/Mol com RF/AAC. Com os atributos híbridos, o BR obteve uma melhoria em torno de 60 Kcal/Mol e com o RF uma melhoria de quase 20 Kcal/Mol. Esses peptídeos foram gradualmente melhorados pelo framework. Em suma, esta análise sugere que SAGAPEP é capaz de convergir adequadamente a busca por peptídeos contra o vírus SARS-CoV-2 usando os parâmetros mencionados anteriormente.



(a) Configuração BR/AAL.

(b) Configuração RF/AAC.

(c) Configuração BR/Híbrido.

(d) Configuração RF/Híbrido.

Figura 15 – Análise de convergência do SAGAPEP em termos do fitness médio dos melhores peptídeo ao longo das gerações de cada execução. (a) demostra a convergência do AG guiado pelo *surrogate* com a configuração BR/AAL. (b) mostra a evolução com o modelo *surrogate* com a configuração RF/AAC. (c) apresenta a evolução com o *surrogate* usando o modelo de regressão BR e o modelo híbrido na extração de atributos (AAL+CKSAAGP). (d) apresenta a evolução com o modelo de regressão RF usando o modelo híbrido na extração de atributos (AAC+AAL).

Em seguida foi avaliado o potencial de descoberta de peptídeos do SAGAPEP avaliando os melhores peptídeos retornados nas dez simulações de cada configuração. Tal aná-

lise foi conduzida por especialistas que modelaram essas sequências peptídicas e avaliaram suas energias de ligação à proteína alvo a partir do software de acoplamento molecular HPEPDOCK.

A Tabela 4 apresenta os valores preditos pelo *surrogate* e os valores de energia de ligação retornados pelos especialistas para os 10 principais peptídeos descobertos pelo SAGAPEP com a configuração BR/AAL. Também é apresentado a respectiva classificação em comparação com os peptídeos do conjunto de dados de treinamento. Desses 10 peptídeos, 8 obtiveram melhores energias de ligação do que todos os peptídeos disponíveis no conjunto de dados utilizado para treinar os modelos *surrogates*, os outros dois também alcançaram resultados consistentes, um classificando-se no top 3 e o outro no top 15. A média de energia de ligação dos peptídeos do conjunto de dados usado no treinamento é de -175.975 Kcal/Mol e o peptídeo com menor avaliação tem uma energia de -230.308. Por outro lado, a média das avaliações dos peptídeos encontrados pelo SAGAPEP usando essa configuração como modelo *surrogate* é de -242.096 Kcal/Mol e o peptídeo com menor avaliação tem uma energia de -272.136 Kcal/Mol. Tais resultados demonstram que essa configuração é capaz de guiar o AG adequadamente na busca de peptídeos com potenciais de inibição do vírus, com apenas 10 execuções o SAGAPEP superou a menor avaliação do conjunto de dados em -41.828 Kcal/Mol, além de uma melhoria considerável na média das avaliações em comparação com o conjunto de dados de treinamento.

Tabela 4 – Peptídeos descobertos pelo SAGAPEP usando a configuração BR/AAL como modelo *surrogate*.

Peptídeos	SAGAPEP	Valor de Referência	Rank
Pep1	-267.138	-272.136	Top1
Pep2	-252.585	-255.732	Top1
Pep3	-252.200	-253.951	Top1
Pep4	-279.509	-247.887	Top1
Pep5	-264.281	-246.508	Top1
Pep6	-276.858	-238.074	Top1
Pep7	-260.981	-236.144	Top1
Pep8	-266.385	-230.651	Top1
Pep9	-266.138	-229.127	Top3
Pep10	-267.173	-210.753	Top15

A Tabela 5 traz os resultados da configuração RF/AAC como modelo *surrogate*. O peptídeo com melhor avaliação encontrado pelo SAGAPEP usando essa configuração como modelo *surrogate* é de -220.910 Kcal/Mol, classificando-se no top 5 em comparação com o conjunto de dados de treinamento; outros dois peptídeos também classificam-se no top 5, dois peptídeos classificam-se no top 10, quatro no top 20 e apenas um ficou fora do top 30. A média das avaliações dos 10 peptídeos encontrados por esse modelo é de -211.687, um melhor desempenho em comparação com a média de -175.975 Kcal/Mol do conjunto de

dados de treinamento. Embora as energias de ligações dos peptídeos encontrados por essa configuração sejam consideráveis, elas possuem energias de ligação com valores superiores aos obtidas pelo modelo BR/AAL, apesar de que ambas configurações obtiveram a mesma avaliação em termos de RSME quando avaliado no módulo avaliação peptídica.

Tabela 5 – Peptídeos descobertos pelo SAGAPEP usando a configuração RF/AAC como modelo *surrogate*.

Peptídeos	SAGAPEP	Valor de Referência	Rank
Pep1	-220.867	-220.910	Top5
Pep2	-220.867	-220.910	Top5
Pep3	-220.867	-219.960	Top5
Pep4	-220.867	-217.902	Top10
Pep5	-220.867	-217.265	Top10
Pep6	-220.867	-209.487	Top20
Pep7	-220.867	-206.067	Top20
Pep8	-220.867	-206.067	Top20
Pep9	-220.867	-205.227	Top20
Pep10	-216.408	-193.078	-

O melhor resultado em termos de RSME no módulo avaliação peptídica foi utilizando um modelo híbrido combinando atributos dos descritores de atributos AAL com KSA-AGP com o modelo de regressão BR. A Tabela 6 apresenta os resultados dos peptídeos descobertos pelo SAGAPEP usando essa configuração. Dos 10 peptídeos, 7 obtiveram energias de ligação melhores do que todos os peptídeos disponíveis no conjunto de dados utilizado para treinar os modelos *surrogates*, os outros três também alcançaram resultados plausível, um classificando-se no top 3, top 5 e top 15. O peptídeo mais bem avaliado descoberto pelo SAGAPEP usando essa configuração tem uma energia de ligação de -274.199 Kcal/Mol, o que superou a melhor avaliação do conjunto de dados de treinamento em -43.891 Kcal/Mol, além de superar as avaliações dos demais peptídeos encontrados pelo SAGAPEP com os outros modelos *surrogates*. A média das avaliações dos 10 peptídeos encontrados por esse modelo é de -246.007 Kcal/Mol, o que supera em quase -71 Kcal/Mol a média do conjunto de dados de treinamento, além de superar aos demais modelos *surrogates* analisados.

A Tabela 7 mostra os resultados dos peptídeos descobertos pelo SAGAPEP usando a configuração RF e o modelo híbrido combinando atributos dos descritores de atributos AAC com AAL. Com essa configuração, o peptídeo encontrado que possui o menor poder de interação tem uma energia de ligação de -237.693 Kcal/Mol, classificando-se no top 1 em comparação com o conjunto de dados de treinamento; outro peptídeo também classifica-se no top 1 e um no top 3, os outros classificam-se no top 5, 10, 15, 20 e apenas um fora do Top 30. A média das avaliações dos 10 peptídeos encontrados por esse modelo é de -218.228 Kcal/Mol, o que é inferior a média usando apenas o extrator de atributos

Tabela 6 – Peptídeos descobertos pelo SAGAPEP usando a configuração BR/Híbrido como modelo *surrogate*.

Peptídeos	SAGAPEP	Valor de Referência	Rank
Pep1	-287.323	-274.199	Top1
Pep2	-277.678	-273.747	Top1
Pep3	-265.931	-267.230	Top1
Pep4	-270.058	-253.016	Top1
Pep5	-281.584	-244.867	Top1
Pep6	-271.588	-240.257	Top1
Pep7	-263.714	-238.504	Top1
Pep8	-268.479	-229.397	Top3
Pep9	-272.124	-224.087	Top5
Pep10	-267.615	-214.764	Top15

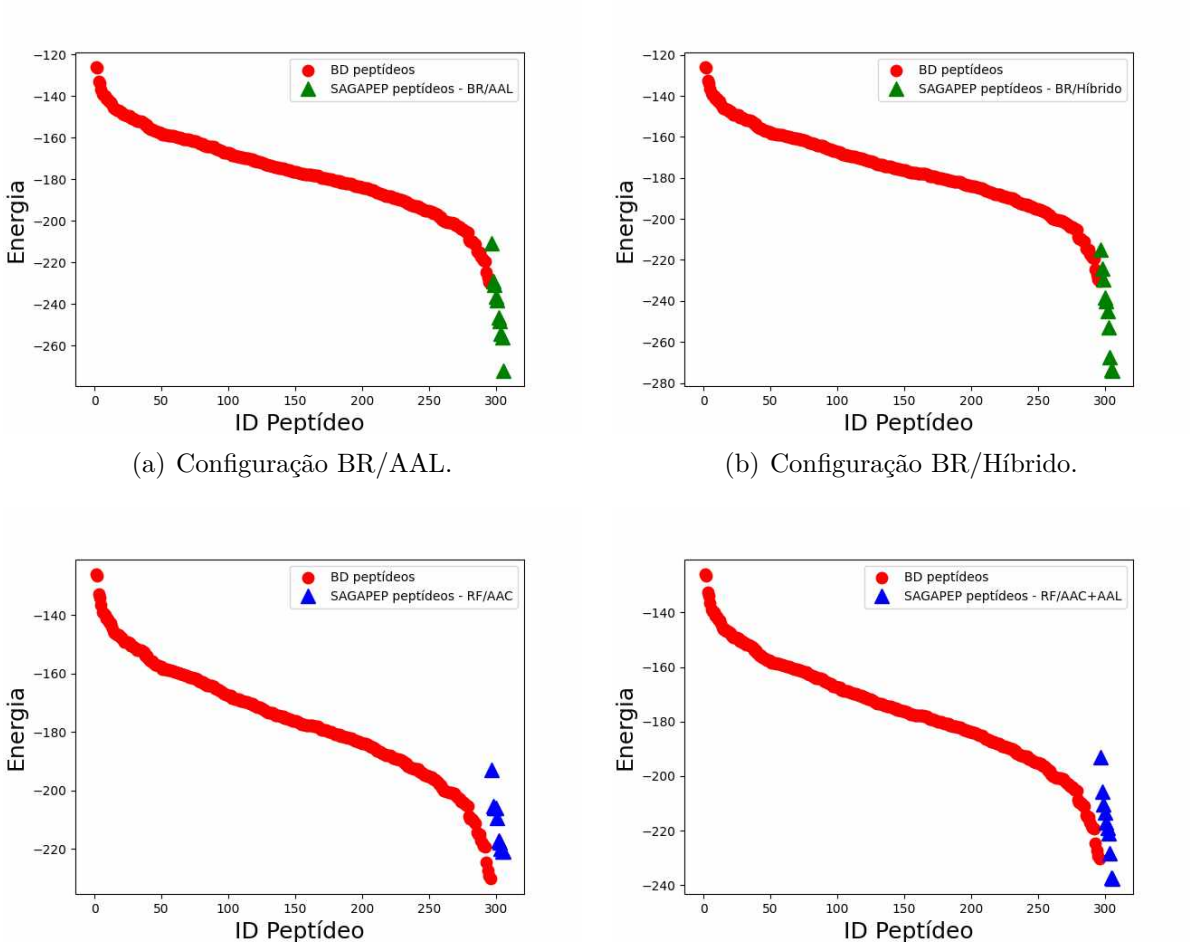
AAC com o modelo de regressão RF.

Tabela 7 – Peptídeos descobertos pelo SAGAPEP usando a configuração RF/Híbrido como modelo *surrogate*.

Peptídeos	SAGAPEP	Valor de Referência	Rank
Pep1	-222.161	-237.693	Top1
Pep2	-215.343	-237.091	Top1
Pep3	-221.487	-228.384	Top3
Pep4	-219.193	-221.074	Top5
Pep5	-217.300	-218.853	Top10
Pep6	-220.581	-216.887	Top10
Pep7	-222.701	-213.257	Top15
Pep8	-222.701	-210.383	Top15
Pep9	-213.909	-205.590	Top20
Pep10	-210.840	-193.063	-

A Figura 16 resume as energias de ligação de todos os peptídeos investigados nessa pesquisa, o que inclui o conjunto de dados de treinamento, bem como os fornecidos em cada configuração avaliada do SAGAPEP. Com exceção da análise RF/AAC, a figura mostra claramente que os peptídeos descobertos usando SAGAPEP têm um maior potencial para inibir a proteína alvo em comparação aos disponíveis em nosso conjunto de dados. De fato, o SAGAPEP em 40 execuções foi capaz de fornecer 17 peptídeos com até 10 resíduos cujas energias de ligações são melhores do que todos os peptídeos no conjunto de dados de treinamento. Isso é particularmente interessante porque pequenos peptídeos são mais baratos, levam menos tempo para sintetizar e simplificam todo o processo de design de peptídeos.

A Figura 17 expõe as energias de ligação dos 50 melhores peptídeos do conjunto de dados de treinamento e as energias dos peptídeos encontrados pelo SAGAPEP. A figura apresenta que o modelo de regressão BR obteve melhor desempenho que o modelo



(a) Configuração BR/AAL.

(b) Configuração BR/Híbrido.

(c) Configuração RF/AAC.

(d) Configuração RF/Híbrido.

Figura 16 – Energias de ligação dos peptídeos do conjunto de dados e os encontrados pelo SAGAPEP. (a) apresenta os peptídeos do conjunto de dados com os encontrados pelo SAGAPEP com a configuração BR/AAL. (b) exibe com os peptídeos encontrados usando a configuração BR/Híbrido. (c) com a configuração RF/AAC e (d) com a configuração RF/Híbrido.

de regressão RF. Dos 20 peptídeos encontrados pelo SAGAPEP usando o modelo de regressão BR na configuração do *surrogate*, 15 classificam-se no top 1 em comparação com o conjunto de dados de treinamento e todos os outros classificam-se no top 15, enquanto que na configuração com o modelo de regressão RF apenas dois estão no top 1 e diversos peptídeos estão fora do top 15. Em relação aos modelos que utilizam apenas um método de extração de atributos e os métodos híbridos, no modelo *surrogate* com a configuração usando o modelo de regressão RF o método híbrido alcançou resultados superiores aos do modelo que utiliza apenas o método de extração de atributos AAC. Já com o modelo de regressão BR os resultados são muito próximos. Dessa forma, as configurações BR/AAL e BR/Híbrido foram as configurações do modelo *surrogate* mais eficientes nos experimentos realizados por essa pesquisa.

Para avaliar empiricamente o tempo de execução do SAGAPEP realizamos simulações

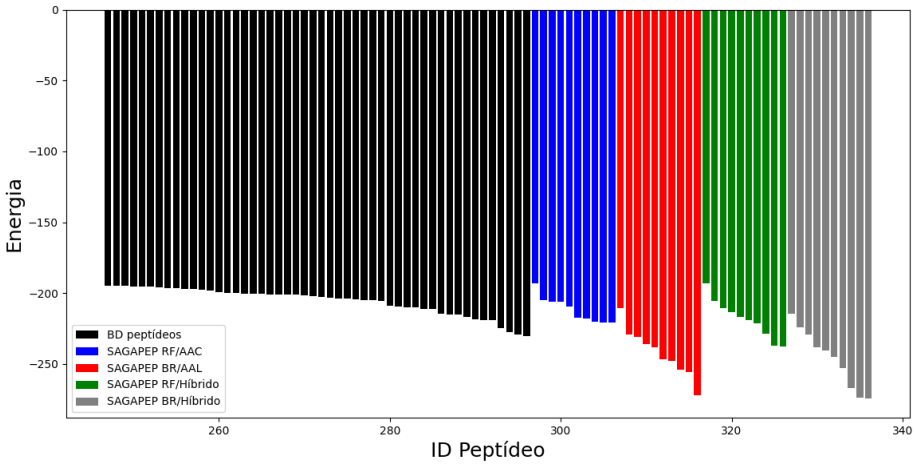


Figura 17 – Energias de ligação dos 50 melhores peptídeos do conjunto de dados de treinamento e dos 40 peptídeos encontrados pelo SAGAPEP e avaliados por especialistas.

considerando as etapas de avaliação peptídica e descoberta de peptídeos usando um computador pessoal com prcessador core i7 e 8 GB de memória RAM. Executamos o módulo avaliação peptídica para predizer a energia de ligação de três peptídeos do conjunto de dados de treinamento: peptídeo com maior quantidade de resíduos; peptídeo com menor quantidade de resíduos; e peptídeo com 16 resíduos (média de resíduos dos peptídeos disponíveis para treinamento). Como resultado, o tempo médio necessário para realizar tais predições foi inferior a um segundo (0.9 segundos). Além disso, também calculamos a média do tempo de execução do módulo de descoberta de peptídeos após 10 execuções. Como resultado, SAGAPEP necessitou em média de menos de dois minutos (108 segundos) para realizar a busca direcionada de peptídeos com potencial de interação com a proteína alvo. Tais resultados, obtidos a partir de um computador pessoal, evidenciam que o framework é capaz de realizar a avaliação e descoberta de peptídeos com reduzido tempo de espera utilizando computadores pessoais.

A Figura 18 apresenta a conformações ligação de alguns desses peptídeos com a proteína Spike do SARS-CoV-2 após a realização do docking molecular usando o software HPEPDOCK. Por fim, ressalta-se que as sequências peptídicas descobertas pelo SAGAPEP não foram divulgadas nessa pesquisa, pois são peptídeos com possibilidades de patentes e transferência para o setor produtivo, atualmente em processo de testagem em bancada.

(a) Pep01 BR/AAL. (b) Pep03 BR/Híbrido. (c) Pep01 RF/Híbrido.

Figura 18 – Exemplos de conformidade de ligação de peptídeos com a proteína Spike do SARS-CoV-2 usando o software HPEPDOCK. (a) mostra a conformidade do Pep01 da configuração BR/AAL. (b) do Pep03 da configuração BR/Híbrido e (c) o Pep01 da configuração RF/Híbrido

Capítulo 5

Conclusão

Este trabalho teve como objetivo principal o desenvolvimento e a aplicação de métodos de otimização bioinspirados assistidos por *surrogates* para a descoberta e seleção de peptídeos com potenciais de inibição à transmissão e disseminação do SARS-CoV-2. O framework desenvolvido foi denominado SAGAPEP, Algoritmo Genético Assistido por *Surrogate* para seleção de peptídeos, e mostrou-se capaz de descobrir peptídeos, bem como prever o poder de sua interação com a proteína Spike do SARSCoV-2. O modelo *surrogate* mostrou-se apto a realizar previsões rápidas e de alta fidelidade da energia de ligação entre um peptídeo e uma proteína alvo. O AG proposto mostrou-se capaz de realizar busca de peptídeos com potencial para inibir a infecção por SARS-CoV-2.

As hipóteses levantadas nessa pesquisa, relacionadas à eficiência de modelos *surrogates* e de otimização bioinspirada respectivamente para avaliação e descoberta de peptídeos, foram confirmadas a partir de evidências obtidas através de avaliações empíricas. O modelo *surrogate* com melhor desempenho alcançou um RMSE de 13.7. Ao mesmo tempo, o modelo de otimização a partir deste surrogate descobriu 10 peptídeos, sendo que 7 deles alcançaram energias de ligação melhores que todos presente na base de dados de treinamento, o que demostra a efetividade do SAGAPEP.

Em relação aos objetivos específicos, destacam-se as seguintes contribuições:

- **Criar representações computacionais que permitem a extração de características em sequências peptídicas**

  Ao todo foram analisados quatro métodos de extração de recursos, a saber: composição de aminoácidos(AAC), ligação dos aminoácidos (AAL), composição de pares de aminoácidos com espaçamento k (CKSNAP) e composição de pares de grupos de aminoácidos com espaçamento k (CKSAAGP). Além disso, foram realizados experimentos usando combinações dois a dois dos métodos de extração de atributos.

- **Treinar modelos *surrogates* capazes de avaliarem de forma rápida e eficiente o valor de interação entre um peptídeo e a proteína Spike do SARS-CoV-2**

Neste trabalho foram conduzidos experimentos que avaliam o desempenho de cinco modelos de aprendizado de máquina como modelo *surrogate* para prever o valor de interação entre um peptídeo e a proteína alvo, usando os métodos de extração de recursos descritos na subseção 3.2.2. Os experimentos mostraram que os modelos *surrogates* são capazes de realizar previsões rápidas e de alta fidelidade da energia de ligação entre um peptídeo e uma proteína alvo.

- ❑ **Desenvolver métodos de otimização bioinspirados para busca de sequências peptídicas de tamanhos variados com potencial de inibição à transmissão e disseminação do SARS-CoV-2**

  Como método de otimização foi desenvolvido um AG capaz de realizar a busca de peptídeos com potencial para inibição do vírus. Além disso, o AG contém um parâmetro denominado *pep_max_size* que permite o controle do número máximo de resíduos presente nos peptídeos encontrados pelo framework.

5.1 Trabalhos Futuros

A presente pesquisa restringiu-se a analisar o SAGAPEP na avaliação e descoberta de peptídeos com potenciais de interação com a proteína Spike do SARS-CoV-2, como trabalho futuro sugere-se a aplicação do framework visando ouras proteínas alvos.

Um das principais limitações encontradas por essa pesquisa está relacionada ao número de amostras presente no conjunto de dados de treinamento dos modelos *surrogates* e a quantidade de resíduos dessas amostras. Como o número de amostras é relativamente pequeno, isso pode prejudicar a avaliação e poder de generalização dos modelos. Além disso, apenas 12 amostras possuem mais de 30 resíduos, o que pode prejudicar o desempenho dos modelos *surrogates* para peptídeos maiores. Desse modo, novas investigações usando conjuntos de dados com maiores quantidades de amostras e com variedades nas quantidades de resíduos dos peptídeos podem melhorar ainda mais a confiabilidade dos modelos *surrogates*.

No contexto de representação, os modelos de extração de atributos baseados nas representações dos aminoácidos usado nessa pesquisa é relativamente simples e eficaz, no entanto essa abordagem pode deixar de obter características relevantes dos peptídeos. Nesse contexto, sugere-se como oportunidade futura a utilização de técnicas de aprendizado profundo usando imagens de peptídeos no treinamento dos modelos *surrogates*.

5.2 Contribuições em Produção Bibliográfica

Esta pesquisa resultou nas seguintes contribuições em termos de produção bibliográfica:

- Silva, E. A. D.; Palmeira, L. S.; Garcia-Júnior, M. A.; Martins, L. G. A.; Andrade, B. S.; Sabino-Silva, R.; Carneiro, M. G. SAGAPEP: A surrogate-assisted genetic algorithm framework to evaluate and discover novel peptides against SARS-CoV-2 virus (submetido). **Expert Systems with Applications**.

- Registro de programa de computador intitulado **SAGAPEP** junto ao Instituto Nacional de Propriedade Industrial, Processo Nº: BR512022002131-5.

Referências

AGRAWAL, P. et al. Anticp 2.0: an updated model for predicting anticancer peptides. **Briefings in Bioinformatics**, v. 22, 2021.

AKINSOLU, M. O. et al. A parallel surrogate model assisted evolutionary algorithm for electromagnetic design optimization. **IEEE Transactions on Emerging Topics in Computational Intelligence**, v. 3, p. 93–105, 2019.

ALONSO, H.; BLIZNYUK, A.; GREADY, J. Combining docking and molecular dynamic simulations in drug design. **Medicinal Research Reviews**, v. 26, p. 531–568, 2006.

BÄCK, T.; FOGEL, D. B.; MICHALEWICZ, Z. **Evolutionary Computation 1: Basic Algorithms and Operators**. [S.l.]: Institute of Physics Publishing, 2000. v. 1.

BARIGYE, S. J. et al. Evolutionary algorithm-based generation of optimum peptide sequences with dengue virus inhibitory activity. **FUTURE MEDICINAL CHEMISTRY**, v. 13, 2021.

BERTONI, A. A. **Avaliação de características e previsão de sucesso de canções populares brasileiras por meio de Aprendizado de Máquina**. Dissertação (Mestrado) — Universidade Federal de Goiás, 2021.

BHASIN, M.; RAGHAVA, G. Classification of nuclear receptors based on amino acid composition and dipeptide composition. **J Biol Chem**, v. 279, p. 23262–23266, 2004.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer Science+Business Media, 2006. v. 1.

BOAS, L. V. et al. Antiviral peptides as promising therapeutic drugs. **Cell Mol Life Sci**, v. 181, p. 3525–3542, 2019.

BOONE, K. et al. Combining genetic algorithm with machine learning strategies for designing potent antimicrobial peptides. **BMC Bioinformatics**, v. 22, p. 239, 2021.

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

BUENO, M. L. d. P. **Heurísticas e algoritmos evolutivos para formulações mono e multiobjetivo do problema do roteamento multicast**. Dissertação (Mestrado) — Universidade Federal de Uberlândia„ 2010.

CAO RUIFEN AMD WANG, M.; BIN, Y.; ZHENG, C. Dlff-acp: prediction of acps based on deep learning and multi-view features fusion. **Peerj**, p. 13, 2021.

CARNEIRO, M. G. **Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural**. 2017. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-01022017-100223/publico/MurilloGuimaraesCarneiro,*evisada.pdf.Acessadoem*25*defev.de*2022.

CASEIRO, A. et al. Mobyle: a new full web bioinformatics framework. **Journal of Proteome Research**, v. 25, p. 3005–3011, 2009.

CASSAR, D. R.; SANTOS, G. G.; ZANOTTO, E. D. Designing optical glasses by machine learning coupled with a genetic algorithm. **Ceramics International**, v. 47, p. 0555–10564, 2021.

CASTRO, L. N. de; ZUBEN, F. J. V. **Recent Developments in Biologically Inspired Computing**. [S.l.]: Idea Group Publishing, 2005.

CHEN, C.; SHEN, Z.-B.; ZOU, X.-Y. Dual-layer wavelet svm for predicting protein structural class via the general form of chou's pseudo amino acid composition. **Sci RepBentham Science Publishers**, v. 19, 2012.

CHEN, Z. et al. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. **PLoS One**, v. 6, 2011.

_____. ilearnplus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. **Nucleic Acids Research**, v. 49, 2020.

CHOU, K. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. **Biochem Biophys Res Commun**, v. 278, p. 477–483, 2000.

CHOU, K. C. Prediction of protein cellular attributes using pseudoamino acid composition. **Proteins: Struct. Funct. Genet**, v. 43, p. 246– 255, 2001.

CHOU, K.-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. **Bioinformatics**, v. 21, p. 10–19, 2001.

CORREIA, G. et al. Airborne route and bad use of ventilation systems as non-negligible factors in sars-cov-2 transmission. **Med Hypotheses**, v. 141, 2020.

CUPERTINO, T. H.; ZHAO, L.; CARNEIRO, M. G. Network-based supervised data classification by using an heuristic of ease of access. **Neurocomputing**, v. 149, p. 86–92, 2015.

DIETTERICH, T. Ensemble methods in machine learning. in: Multiple classifier systems. **Springer**, v. 1857, p. 1–15, 2000.

DING, S.; LIANG, T. J. Is sars-cov-2 also an enteric pathogen with potential fecal–oral transmission? a covid-19 virological and clinical review. **Gastroenterology**, v. 156, p. 53–61, 2020.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. New York: Wiley, 2001. ISBN 978-0-471-05669-0.

FLETCHER, T. **Support Vector Machines Explained**. 2008. Disponível em: urlhttps://www.csd.uwo.ca/ xling/cs860/papers/SVM$_E$*xplained.pdf*.*Acessadoem*14*dejan.de*2022.

GOLDBERG, D.; HOLLAND, J. Genetic algorithms and machine learning. **Machine Learning**, v. 3, 1988.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization, and Machine Learning**. [S.l.]: Addison-Wesley, 1989.

GUPTA, A.; BHAKTA, S. An integrated surrogate model for screening of drugs against mycobacterium tuberculosis. **Journal of Antimicrobial Chemotherapy**, v. 67, p. 1380–1391, 2012.

HAN, X. et al. Improving protein solubility and activity by introducing small peptide tags designed with machine learning models. **Metabolic Engineering Communications**, v. 11, 2020.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. [S.l.]: Springer, 2014. v. 2.

HAYKIN, S. **Redes Neurais: Princípios e Práticas**. [S.l.]: Bookman, 2001. v. 2.

HU, B. et al. Characteristics of sars-cov-2 and covid-19. **Nat Rev Microbiol**, v. 19, p. 141–154, 2021.

HUANG, S.; ZOU, X. An iterative knowledge-based scoring function for protein-rotein recognition. **Proteins**, v. 72, p. 557–579, 2008.

JOHANSSON-ÅKHE, I.; MIRABELLO, C.; WALLNER, B. Predicting protein-peptide interaction sites using distant protein complexes as structural templates. **Biopolymers**, v. 9, 2018.

JU, Z.; HE, J. Prediction of lysine propionylation sites using biased svm and incorporating four different sequence features into chou's pseaac. **Journal of Molecular Graphics and Modelling**, v. 76, p. 356–363, 2017.

JU, Z.; WANG, S. Prediction of citrullination sites by incorporating k-spaced amino acid pairs into chou's general pseudo amino acid composition. **Gene**, v. 664, p. 356–363, 2018.

KABRA, R.; SINGH, S. Evolutionary artificial intelligence based peptide discoveries for effective covid-19 therapeutics. **Biochimica et Biophysica Acta**, v. 1867, 2021.

KAO, H. et al. Succsite: Incorporating amino acid composition and informative k-spaced amino acid pairs to identify protein succinylation sites. **Genomics, Proteomics Bioinformatics**, v. 18, 2020.

KITCHEN, D. et al. Docking and scoring in virtual screening for drug discovery: methods and applications. **Nat. Rev. Drug Discov**, v. 3, p. 935–949, 2004.

KRAMER, O. K-nearest neighbors. **Intelligent Systems Reference Library**, v. 51, p. 13–23, 2013.

KUMAR, R.; CHAUDHARY, K.; CHAUHAN, J. e. a. S. An in silico platform for predicting, screening and designing of antihypertensive peptides. **Sci Rep**, v. 12512, 2015.

LINDEN, R. **Algoritmos Genéticos.** [S.l.]: Ciência Moderna Ltda., 2012. v. 3.

MACKAY, D. J. Bayesian interpolation. **Computation and Neural Systems**, v. 4, p. 139–74, 1992.

MANAVALAN, B.; BASITH, S.; LEE, G. Comparative analysis of machine learning-based approaches for identifying therapeutic peptides targeting sars-cov-2. **Briefings in Bioinformatics**, v. 23, 2022.

MARQUS, S.; PIROGOVA, E.; PIVA, T. J. Evaluation of the use of therapeutic peptides for cancer treatment. **Journal of Biomedical Science**, v. 24, 2017.

MENG, X. et al. Molecular docking: A powerful approach for structure-based drug discovery. **Curr Comput Aided Drug**, v. 26, p. 146–57, 2011.

MOLNAR, C. **Interpretable Machine Learning: A Guide for Making Black Box Models Explainable**. [S.l.]: Lulu, 2018.

MORRIS, G.; LIM-WILBY, M. Molecular docking. **PKukol A. (eds) Molecular Modeling of Proteins. Methods Molecular Biology™**, v. 443, 2008.

NELSON, D. L.; COX, M. M. **Lehninger Principles of Biochemistry**. [S.l.]: Macmillan Higher Education Houndmills., 2017. v. 7.

NÉRON, B. et al. Salivary proteome and peptidome profiling in type 1 diabetes mellitus using a quantitative approach. **Bioinformatics**, v. 12, p. 1700–1709, 2013.

NIWA, T. et al. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of escherichia coli proteins. **Proc. Natl. Acad. Sci.**, v. 106, p. 4201 – 4206, Niwaa2009.

OLIVEIRA, T. L. et al. Pathophysiology of sars-cov-2 in lung of diabetic patients. **Frontiers in Physiology**, v. 11, p. 139–74, 2020.

PAL, S. K.; MITRA, S. Multilayer perceptron, fuzzy sets, and classification. **in IEEE Transactions on Neural Networks**, v. 3, p. 683–697, 1992.

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

QURESHI, A.; TANDON, H.; KUMAR, M. Avp-ic50pred: Multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (ic50). **Biopolymers**, v. 6, 2015.

RUSSELL, S. J.; NORVIG, P. **Inteligência artificial**. [S.l.]: Elsevier, 2013. v. 7.

SILVA, E. A. D. et al. Sagapep: A surrogate-assisted genetic algorithm framework to evaluate and discover novel peptides against sars-cov-2 virus. **Expert Systems with Applications**, 2022.

SMOLA, A.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and Computing**, v. 14, p. 99–222, 2004.

TIPPING, M. E. Sparse bayesian learning and the relevance vector machine. **Journal of Machine Learning Research**, v. 1, p. 211–244, 2001.

ULLAH, Z. et al. A comparative study of machine learning methods for bio-oil yield prediction – a genetic algorithm-based features selection. **Bioresource Technology**, v. 335, p. 125292, 2021.

VIELHABEN, M. W. J.; WEICKEN, E.; STRODTHOFF, N. **Predicting the Binding of SARS-CoV-2 Peptides to the Major Histocompatibility Complex with Recurrent Neural Networks**. arXiv, 2021. Disponível em: <https://arxiv.org/abs/2104.08237>.

WEI, L. et al. Acpred-fl: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. **Bioinformatics**, v. 34, p. 4007–4016, 2018.

WHITLEY, D. A. A genetic algorithm tutorial. **Stat Comput**, v. 4, p. 65–85, 1994.

WORLD, H. O. **Coronavirus disease (COVID-19)**. 2022. Disponível em: urlhttps://www.who.int/eportuguese/countries/bra/pt/. Acessado em 04 de agosto de 2022.

YANG, J. et al. Molecular interaction and inhibition of sars-cov-2 binding to the ace2 receptor. **Nature Communications**, v. 11, 2020.

YANG, X.-S. **Nature-Inspired Optimization Algorithms**. [S.l.]: Elsevier, 2014. v. 1.

YOSHIDA, M. et al. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. **Chem**, v. 4, p. 533–543, 2018.

ZHAI, Y. et al. Identifying antioxidant proteins by using amino acid composition and protein-protein interactions. **Front Cell Dev Biol**, v. 8, 2020.

ZHANG, X. et al. Multi-fidelity deep neural network surrogate model for aerodynamic shape optimization. **Computer Methods in Applied Mechanics and Engineering**, v. 373, p. 113485, 2021.

ZHOU, P. et al. Hpepdock: a web server for blind peptide–protein docking based on a hierarchical algorithm. **Nucleic Acids Res**, v. 46, p. 443–445, 2018.

Apêndices

APÊNDICE **A**

Resultados com Outros Valores de *k*

Neste apêndice são apresentados os experimentos realizados para a definição do valor de k para os métodos de extração de de recursos CKSNAP e CKSAAGP. A Tabela 8 mostra os resultados obtidos para os valores de k igual a 1, 2, 3, 4 e 5 para o método de extração de recurso CKSNAP. Os resultados são apresentados em termos de RMSE (média e desvio padrão) e mostram que quatro dos cinco modelos de regressão obtiveram melhores resultados usando $k = 5$, desse modo esse valor foi selecionado como parâmetro para o extrator de recurso CKSNAP.

Tabela 8 – Desempenho dos modelos *surrogates* em termos de RMSE (média e desvio padrão) com o CKSNAP usando diferentes valores de k.

CKSNAP	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
BR	17.5 ± 3.1	15.8 ± 2.3	15.4 ± 2.2	14.67 ± 1.9	14.5 ± 1.9
RF	16.5 ± 2.0	15.6 ± 2.0	15.5 ± 2.0	15.3 ± 1.7	15.0 ± 1.9
KNN	21.7 ± 2.6	21.27 ± 2.4	20.8 ± 2.3	20.5 ± 2.3	20.7 ± 2.2
SVM	15.6 ± 2.0	15.3 ± 2.0	15.0 ± 1.8	14.8 ± 2.0	14.9 ± 2.2
MLP	19.2 ± 2.0	19.2 ± 2.2	19.2 ± 2.4	19.2 ± 2.1	19.2 ± 2.2

A Tabela 9 traz os resultados dos experimentos usando diversos valores de k para o método de extração de recurso CKSAAGP. De modo geral, os resultados não tiveram grandes variações, duas das cinco técnicas obtiveram os melhores resultados com $k = 1$, dessa forma esse valor foi selecionado para configuração do extrator de recurso CKSAAGP.

Tabela 9 – Desempenho dos modelos *surrogates* em termos de RMSE (média e desvio padrão) com o CKSAAGP usando diferentes valores de k.

CKSAAGP	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
BR	15.8 ± 1.9	16.0 ± 2.2	16.1 ± 2.2	16.1 ± 1.9	16.1 ± 2.1
RF	14.3 ± 1.8	14.4 ± 2.0	15.1 ± 2.0	14.6 ± 1.7	14.5 ± 1.8
KNN	16.9 ± 2.4	16.9 ± 2.4	17.2 ± 2.1	16.5 ± 1.7	16.5 ± 2.0
SVM	15.3 ± 2.0	15.2 ± 2.1	15.1 ± 2.1	14.8 ± 1.7	15.1 ± 1.9
MLP	19.2 ± 2.2	19.2 ± 2.6	19.2 ± 2.4	19.2 ± 2.3	19.2 ± 2.2