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

Elias A. D. Silva [1,2]**,Luiz G. A. Martins**[1]**, Murillo G. Carneiro**[1]

[1]Faculdade de Computação – Universidade Federal de Uberlândia(UFU)
Uberlândia – MG – Brasil

[2]Instituto Federal de Educação, Ciência e Tecnologia de Rondônia (IFRO)
Ji-Paraná – RO – Brasil

elias.silva@ifro.edu.br, lgamartins@ufu.br, mgcarneiro@ufu.br

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.*

1. Problema

A descoberta de peptídeo que possa interagir com uma das proteína do SARS-CoV-2 se caracteriza como uma estratégia promissora no sentido da mitigação da transmissão entre hospedeiros do vírus. Os métodos experimentais utilizados para análise de estruturas complexas peptídeo-proteína exigem altos custos e um tempo considerável, dessa forma a modelagem computacional por docking molecular tem desempenhado um papel importante nesse tipo de análise. Ainda assim, essa abordagem necessita que proteína alvo e as moléculas estejam em uma estrutura tridimensional (3D) [Morris and 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. O servidor HPEPDOCK [Huang and Zou 2008], por exemplo, 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 [Zhou et al. 2018]. Além disso, há uma grande quantidade de peptídeos que podem ser formados pelos 20 aminoácidos frequentemente encontrados como constituintes de proteínas, com exatamente 10 resíduos é possível formar $\cong 20^{10}$ peptídeos, no entanto peptídeos normalmente são compostos de 2 a 50 resíduos [Nelson and Cox 2017].

Neste trabalho apresentamos o 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. O SAGAPEP é composto por dois módulos principais: avaliação de peptídeos e descoberta de peptídeos. 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. Para a descoberta de peptídeos, modelamos uma técnica de otimização bioinspirada que adota operadores genéticos baseados na teoria da evolução, denominada Algoritmo Genético (AG), para executar uma busca iterativa de peptídeos com alto potencial para inibir a infecção por SARS-CoV-2.

Dessa forma, A presente pesquisa teve 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 SARS-CoV-2. As contribuições específicas dessa pesquisa são listadas a seguir:

- Projeto de representações computacionais que permitem a extração de atributos em sequências peptídicas;
- Treinamento de 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;
- Desenvolvimento de métodos de otimização bioinspirados para busca de sequências peptídicas de tamanhos variados com potencial de inibição à transmissão e disseminação do SARS-CoV-2.

2. Descrição do Framework SAGAPEP

O desenvolvimento do SAGAPEP foi dividido em dois grandes módulos: avaliação peptídica e descoberta de peptídeos. O módulo avaliação peptídica tem com finalidade prever a energia de ligação entre um peptídeo e a proteína alvo e o módulo descoberta de peptídeos consiste em um algoritmo genético assistido por *surrogate* para a descoberta de peptídeos com potencial inibidor de vírus.

2.1. Avaliação de Peptídeo

A seguir, detalhamos as principais etapas envolvidas no aprendizado de nosso modelo *surrogate*: conjunto de dados, métodos de extração de recursos e treinamento de modelo *surrogate*.

2.1.1. 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). 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 do SARS-CoV-2. Quanto menor esse valor, melhor é a interação do peptídeo contra o alvo.

2.1.2. Métodos de extração de recursos

Os métodos de extração de recursos têm 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: 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) [Chen et al. 2020].

2.1.3. Treinamento dos Modelos *Surrogates*

Como modelos *surrogates*, foram treinados e analisados os algoritmos de aprendizado de máquina: florestas aleatórias (RF), regressão bayesiana (BR), máquina de vetor de suporte (SVM), K-vizinhos mais próximos (KNN) e perceptron multicamadas (MLP). 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 obtidos no acoplamento molecular.

2.2. Descoberta de Peptídeo

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 et al. 2018]. O modelo *surrogate* treinado no módulo anterior é utilizado para avaliar a aptidão de cada indivíduo, a cada avaliação um vetor de atributos é calculado utilizando o método de extração de atributos selecionado durante a fase de treinamento do *surrogate*, e então é inserido como entrada para o modelo *surrogate* que retorna a energia de ligação prevista para a sequência contra a proteína alvo.

3. Experimentos e Resultados

Nesta seção 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.

3.1. Avaliação Peptídica com *Surrogate*

A métrica utilizada para avaliação dos modelos *surrogates* é a validação cruzada repetida com partições e 10 repetições, o que resulta em um total de 100 simulações

[Pedregosa et al. 2011]. A Tabela 1 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, os 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.

Tabela 1. 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	$\mathbf{14.1 \pm 2.6}$	14.5 ± 1.9	16.0 ± 1.9
RF	$\mathbf{14.1 \pm 2.0}$	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

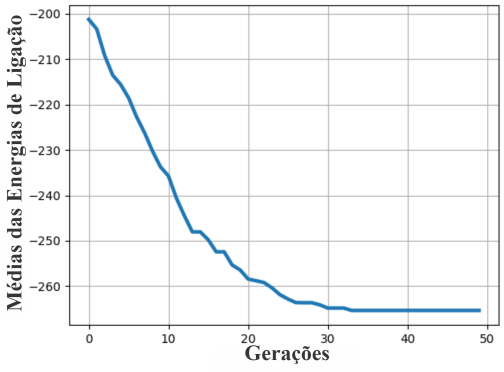
3.2. Descoberta de Peptídeos com SAGAPEP

Para identificar peptídeos com alto potencial de interação com a proteína Spike SARS-CoV-2, realizamos uma simulação de 10 execuções para o modelo *surrogates* com a configuração o BR/AAL. O número máximo de resíduos nos peptídeos foi definido como pep_max_size = 10.

A primeira análise teve como objetivo avaliar a convergência do nosso AG em relação aos parâmetros selecionados. A Figura 1 demonstra que SAGAPEP é capaz de evoluir para melhores peptídeos candidatos ao longo de gerações. Na figura, foi calculada a média do melhor peptídeo por geração considerando as 10 execuções de cada configuração simuladas com nosso AG. Até a 20ª geração houve uma melhoria de aproximadamente 60 Kcal/Mol com BR/AAL.

Em seguida foi aferido o potencial de descoberta de peptídeos do SAGAPEP avaliando os melhores peptídeos retornados nas dez simulações da configuração. Tal análise 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 2 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, o valor de referência é o valor obtido pelos especialistas. Também é apresentada 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. 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



(a) Configuração BR/AAL.

Figura 1. Análise de convergência do SAGAPEP em termos do fitness médio dos melhores peptídeo ao longo das gerações de cada execução.

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.

Tabela 2. Peptídeos descobertos pelo SAGAPEP usando a configuração BR/AAL.

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

O 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 com prcessador core i7 e 8 GB de memória RAM, evidenciam que o framework é capaz de realizar a avaliação e descoberta de peptídeos com reduzido tempo de espera utilizando computadores pessoais. 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.

4. Considerações Finais

A presente pesquisa teve como objetivo a utilização de modelos *surrogates* para avaliação de interação entre peptídeo e a proteína Spike do SARS-CoV-2, além da aplicação de métodos de otimização para busca de peptídeos. Os modelos de aprendizado de

máquina usados como modelos surrogates mostraram-se capazes de predizerem a energia de ligação com baixas taxas de erros. O algoritmo genético mostrou-se eficaz na busca por peptídeos que possam interagir com a proteína Spike, como poucas execuções conseguimos superar a melhor energia de ligação presente na base de dados de treinamento em mais de -40 Kcal/Mol. Todas as simulações foram configurdas com o número máximo de resíduos como 10 (pep max size = 10), em trabalhos futuros será analisado com peptídeos com maiores quantidades de resíduos.

Como contribuições em termos de produção bibliográfica, foi realizado o registro de programa de computador intitulado SAGAPEP junto ao Instituto Nacional de Propriedade Industrial, processo Nº: BR512022002131-5. Também foi apresentado o trabalho intitulado A high throughput screening tool for discovering peptide bioactivity no *International Society for Computational Biology*, 2022. Além disso, o artigo "SAGAPEP: A surrogate-assisted genetic algorithm framework to evaluate and discover novel peptides against SARS-CoV-2 virus"foi escrito e submetido para periódico de relevância (Qualis A1).

Referências

Caseiro, A., Ferreira, R., Padrão, A., Quintaneiro, C., Pereira, A., and Marinheiro, R. (2009). Mobyle: a new full web bioinformatics framework. *Journal of Proteome Research*, 25:3005–3011.

Chen, Z., Zhao, P., Li, C., Li, F., Xiang, D., Chen, Y.-Z., Akutsu, T., Daly, R. J., Webb, G. I., Zhao, Q., Kurgan, L., and Song, J. (2020). ilearnplus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Research*, 49.

Huang, S. and Zou, X. (2008). An iterative knowledge-based scoring function for protein-rotein recognition. *Proteins*, 72:557–579.

Johansson-Åkhe, I., Mirabello, C., and Wallner, B. (2018). Predicting protein-peptide interaction sites using distant protein complexes as structural templates. *Biopolymers*, 9.

Morris, G. and Lim-Wilby, M. (2008). Molecular docking. *PKukol A. (eds) Molecular Modeling of Proteins. Methods Molecular Biology™*, 443.

Nelson, D. L. and Cox, M. M. (2017). *Lehninger Principles of Biochemistry*, volume 7. Macmillan Higher Education Houndmills.

Néron, B., Ménager, H., Maufrais, C., Joly, N., Maupetit, J., Letort, S., Carrere, S., Tuffery, P., and Letondal, C. (2013). Salivary proteome and peptidome profiling in type 1 diabetes mellitus using a quantitative approach. *Bioinformatics*, 12:1700–1709.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Zhou, P., Jin, B., Li, H., and Huang, S. (2018). Hpepdock: a web server for blind peptide–protein docking based on a hierarchical algorithm. *Nucleic Acids Res*, 46:443–445.