aCSM-all como validação de complexos de proteínas gerados pelo AlphaFold

Taís Christofani[1], **Raquel C. de Melo-Minardi**[1]

[1]Departamento de Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, MG – Brazil

{tais.christofani,raquelcm}@dcc.ufmg.br

Abstract. *The protein folding problem is one of the greatest open challenges in biology and consists of determining the molecular structure of a protein based only on its genetic sequence. This determination is extremely important, since the biological function of a protein is intrinsically linked to its structure. In 2018, AlphaFold artificial intelligence emerged to predict protein structures, obtaining such good results that it became widely used by the scientific community. When it comes to protein interactions, AlphaFold still faces challenges and it is difficult to say which protein complexes modeled by it actually have a biological function. Thus, the present work proposed to create a classifier for complex structures generated by AlphaFold to identify the structures as native or not. Using biological graph signatures generated by the aCSM-all method as target function for different machine learning techniques, an AUC parameter of 0.610 was obtained with the kNN algorithm, slightly better than a random classifier. More studies will be needed to say whether signatures serve as a parameter for the classification in question or not.*

Resumo. *O problema do enovelamento de proteínas é um dos grandes desafios em aberto da biologia e consiste em determinar a estrutura molecular de uma proteína com base apenas na sua sequência genética. Essa determinação é extremamente importante, já que a função biológica de uma proteína está intrinsecamente ligada a sua estrutura. Em 2018, surgiu a inteligência artificial AlphaFold para a previsão de estruturas de proteínas, obtendo resultados tão bom que passou a ser amplamente utilizada pela comunidade científica. Ao que diz respeito à interação de proteínas, o AlphaFold ainda enfrenta desafios e é difícil dizer que complexos protéicos modelados por ele possuem de fato função biológica. Assim, o presente trabalho se propôs a criar um classificador para estruturas de complexos geradas pelo AlphaFold para identificar as estruturas como nativas ou não. Utilizando as assinaturas de grafos biológicos geradas pelo método aCSM-all como função alvo de diferentes técnicas de aprendizado de máquina, foi obtido um parâmetro AUC de 0,610 com o algoritmo kNN, um pouco melhor que um classificador aleatório. Mais estudos serão necessários para dizer se as assinaturas servem como parâmetro para a classificação em questão ou não.*

1. Capítulo introdutório

Proteínas são compostos orgânicos formados por aminoácidos que estão presentes em todos os seres vivos e desempenham inúmeras funções nos processos celulares desses

organismos. A função de uma proteína está intrinsecamente ligada à sua estrutura molecular, sendo sua determinação de grande valia para a elucidação de diversos processos biológicos. Apesar de sua importância, poucas proteínas já foram resolvidas experimentalmente, já que esse é um processo custoso e demorado. Não à toa, em janeiro de 2023, o número de proteínas com sequência de aminoácidos conhecida era mais de mil vezes maior que a de proteínas com estruturas resolvidas [Bertoline et al. 2023]. Dessa forma, um dos grandes desafios em aberto da biologia é prever a estrutura tridimensional de uma proteína com base no seu código genético.

Para resolver esse desafio, conhecido como o problema de enovelamento de proteínas, a DeepMind treinou uma inteligência artificial que apresentou resultados surpreendentes ao participar da edição de 2021 da competição CASP (Critical Assessment of protein Structure Prediction). A inteligência em questão, chamada AlphaFold, utiliza redes neurais que são alimentadas pela base de dados Protein Data Bank (PDB), contendo todas as proteínas cujas estruturas já foram resolvidas, para aprender seus padrões e prever a estrutura de outras proteínas. Desde sua introdução, devido à sua acurácia e velocidade, o AlphaFold se tornou uma ferramenta-chave entre os cientistas e vem sido utilizada e aprimorada para diversas áreas.

Um dos campos que têm se beneficiado da existência do AlphaFold é o de estudo de interações entre proteínas. As interações proteína-proteína ocorrem entre duas ou mais proteínas que se associam de forma não-covalente formando um complexo proteico. O entendimento dessas interações está envolvido na maioria dos processos biológicos mais intrincados, sendo que pouquíssimos complexos de proteínas presentes nos seres humanos têm sua estrutura experimental elucidada. Embora versões aprimoradas do AlphaFold para essa área, o AlphaFold2 e o AlphaFold-Multimer, façam predições dessas estruturas com uma acurácia melhor que os demais métodos existentes, essa área ainda apresenta desafios. Dois estudos recentes, usando o AlphaFold e o AlphaFold2 mostraram que ele gerou modelos com boa qualidade em 43% [Yin et al. 2022] e 63% [Bryant et al. 2022] das estruturas avaliadas, respectivamente. Assim, se faz interessante a existência de métodos que determinem com confiança se as estruturas ainda não resolvidas experimentalmente, preditas por essas tecnologias, são nativas biologicamente ou não, ou seja, se suas estruturas são de fato interagentes e corretas.

Devido a esses fatores, é proposta a criação de um classificador de complexos proteicos gerados pelo AlphaFold2 tendo como *input* proteínas de sequências conhecidas. Para isso, pretende-se expandir o uso de uma assinatura estrutural de proteínas, criada por Pires *et al.* [Pires et al. 2011b], inicialmente desenvolvida para relacionar a estrutura de proteínas com sua função biológica. O objetivo então é usar essa assinatura como parâmetro de algoritmos de aprendizado de máquina que permita classificar as interações proteína-proteína como nativas biologicamente ou não.

2. Capítulo referencial

Existem diversos métodos para mensurar a qualidade dos complexos de proteínas geradas por ferramentas de *docking* que realizam um ranking para achar as estruturas mais promissoras dentre as várias geradas por uma ou mais ferramentas [Moal et al. 2013]. Entretanto, elas não se propõem a fazer uma medição absoluta da qualidade de uma estrutura. Em 2016, foi desenvolvido o DockQ [Basu and Wallner 2016a] para avaliar

um complexo proteico com uma pontuação contínua, baseado no protocolo de avaliação CAPRI [Lensink et al. 2007], utilizado na competição CASP mencionada anteriormente. Assim como o CAPRI, que divide as estruturas em quatro classificações apenas, o DockQ é calculado usando uma combinação de inúmeros critérios, incluindo a energia de dessolvatação, a geometria da estrutura e a consistência dos contatos entre as proteínas. No artigo original, os autores dizem que essa pontuação pode servir de função alvo em algoritmos de aprendizado de máquina, porém só foi achado um estudo com aprendizado de máquina focado em encontrar complexos proteicos nativos [Basu and Wallner 2016b]. O classificador proposto no presente trabalho se pretende expandir o desenvolvimento de medidores de qualidade, se diferenciando por usar como função alvo um parâmetro desenvolvido por Pires *et al.*: o aCSM-all, derivado do Cutoff Scanning Matrix (CSM), um modelo para geração de assinaturas para grafos biológicos.

O CSM se baseia apenas nas distâncias entre resíduos para propor uma assinatura para estruturas de proteínas. Ela serviria então como uma forma de identificar o enovelamento da proteína e a natureza das interações que ela pode estabelecer com outras proteínas e ligantes [Pires et al. 2011b]. A assinatura é obtida como uma distribuição cumulativa de contato, medido como a quantidade de átomos de carbonos alfa a uma determinada distância, ou *cutoff*, uns dos outros. A ideia é que proteínas com diferentes tipo de enovelamento apresentariam diferentes distribuições dessas distâncias. Diferentes modificações desse método foram desenvolvidos e se mostraram útil para as mais variadas aplicações, como predição de efeitos de mutação em proteínas [Pires et al. 2013a], previsão da farmacocinética e toxicidade de novas drogas [Pires et al. 2015] e identificação inibidores de interações proteína-proteína [Rodrigues et al. 2021].

O aCSM-all [Pires et al. 2013c] é uma extensão do CSM em que: (i) são considerados todos os átomos na distribuição cumulativa de contato, e não só os carbonos alfas; (ii) além da quantidade de átomos pra cada distância, são também calculados o número de pares de átomos que se encaixam em cada combinação dois a dois de oito categorias (hidrofóbico, positivo, negativo, aceptor, doador, aromático, sulfúrico e neutro), resultando em 36 combinações por *cuttof*.

3. Capítulo de contribuição

O projeto contou com a contribuição de um aluno da disciplina de Bioinformática do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais, já que foi apresentado também como projeto final em dupla da matéria.

Inicialmente, foi realizada a seleção dos complexos a serem utilizados para o cálculo das assinaturas e, posteriormente, utilizadas como base de dados para o treinamento dos classificadores. Os complexos foram retirados do banco de dados PDB, em que se utilizou a ferramenta de busca avançada para filtrar apenas os complexos desejados, com duas proteínas e caracterizadas por difração de raios-x. Foram coletadas 100 complexos dessa forma, rotulados como nativos. Os complexos possuem a cadeia A e B, sendo que para a geração dos complexos não-nativos, para cada complexo, a cadeia B foi removida e adicionada a cadeia B de algum outro complexo nativo.

Em seguida, foi iniciado o processo de geração das estruturas pelo Google colab AlphaFold2. Para isso, bastou colocar, para cada um dos 200 complexos selecionados, as sequências das suas cadeias A e B como input do documento. O AlphaFold2, então,

encontrou a estrutura dos complexos em formato *.pdb*, além de outros arquivos para diferentes finalidades. Foi utilizada a estrutura *rank001*, que seria a melhor previsão dentre as geradas.

De posse das estruturas geradas pelo AlphaFold2, foi utilizada a biblioteca Signa para geração das assinaturas dessas estruturas. A biblioteca é *open source* e foi criada por um ex-aluno do grupo de pesquisa da professora Raquel Minardi, do Departamento de Ciência da Computação da Univerdisdade Federal de Minas Gerais. A biblioteca gerou um arquivo *.csv* com a assinatura de todas as estruturas.

Com o auxílio da ferramenta Orange Data Mining [Demšar et al. 2013], esse arquivo foi selecionado como nossa base de dados, com cada valor do vetor das assinaturas sendo uma *feature* do nosso modelo de classificação entre complexos nativos e não-nativos. Então, foi feita uma validação cruzada de 20 *folds* com as seguintes técnicas para classificação: kNN, SVM, Regressão Logística, Naive Bayes e Random Forest. Para todos eles foram utilizados os parâmetros padrão dados pela própria ferramenta Orange.

4. Capítulo de fechamento

As principais métricas obtidas estão disponibilizadas na tabela 1 e figura 1. Cada uma nos informa algo sobre a predição e pode ser melhor que outra para avaliar a performance do classificador. As matrizes confusão demonstram o fato das nossas classes estarem balanceadas, uma vez que a proporção de acertos entre as classes nativa e não-nativa é semelhante, do mesmo modo que os erros. Essa proporção sugere que o modelo não está favorecendo uma classe sobre a outra, o que é positivo no nosso caso, em que não há uma classe dominante.

A acurácia fornece um desempenho geral do modelo, ou quantos acertos ele teve em todas as classificações, podendo ser enganosa quando uma classificação é mais importante que a outra. A precisão mede a acurácia do modelo para a classe positiva, mais usada em casos em que os falsos positivos são considerados mais problemáticos que os falsos negativos. O *recall* já diz respeito apenas a como as amostras positivas são classificadas, avaliando a sensibilidade do classificador, ou quantas amostras positivas foram classificadas como positivas com relação a quantas era positivas de fato, usado em situações em que os falsos negativos são vistos como mais problemáticos que os falsos positivos. O *F1 score* é apenas uma média harmônica entre *recall* e precisão. Como nossa base de dados é balanceada e a determinação das duas classes é igualmente importante, não faz tanta diferença qual a seleção da classe positiva e da negativa, e também as quatro métricas são boas para avaliar o modelo. Além disso, como visto na tabela, eles não diferem muito entre si para cada classificador.

	AUC	Acurácia	F1	Precisão	*Recall*
kNN	0,610	0,586	0,595	0,597	0,596
Naive Bayes	0,565	0,545	0,545	0,545	0,545
SVM	0,377	0,545	0,543	0,546	0,545
Random Forest	0,521	0,515	0,515	0,515	0,515
Regressão Logísitca	0,480	0,479	0,479	0,480	0,480

Tabela 1. Métricas obtidas para os classificadores avaliados

Figura 1

Matrizes confusão para os classificadores avaliados

kNN

	Não nativo	Nativo
Não nativo	58,7%	39,3%
Nativo	41,3%	60,7%

Naïve Bayes

	Não nativo	Nativo
Não nativo	54,5%	45,4%
Nativo	45,5%	54,6%

SVM

	Não nativo	Nativo
Não nativo	54%	44,7%
Nativo	46%	55,3%

Random Forest

	Não nativo	Nativo
Não nativo	51,5%	48,5%
Nativo	48,5%	51,5%

Regressão Logística

	Não nativo	Nativo
Não nativo	48,1%	52,1%
Nativo	51,9%	47,9%

Já o AUC (*Area Under Roc curve*) é uma das métricas mais utilizadas para avaliar a performance de algoritmos de aprendizado de máquina, já que a curva ROC resume o *trade-off* entre sensibilidade e especificidade, medindo quanto o modelo é capaz de discernir entre as classes.

As métricas analisadas variam de 0 a 1, sendo 1 um classificador perfeito e 0,5 um classificador totalmente aleatório. Os nossos dados nos mostram que o melhor modelo foi o kNN, que apesar de não ter tido um ótimo desempenho, conseguiu distinguir entre as duas classes de complexo de forma melhor que aleatória, o que indica que há espaço para mudanças no método proposto de forma a aprimorar esse valor.

Apesar de se esperar uma performance melhor do SVM, devido à sua habilidade de lidar com problemas de alta dimensionalidade e ter menos risco de *overfitting*, o classificador kNN já se mostrou o melhor também para outros estudos usando assinaturas, sendo utilizado por padrão em demais pesquisas com elas [Pires et al. 2013b][Pires et al. 2011a]. Não se sabe por que isso acontece, mas pode ser pelo fato de que ele não faz suposições sobre a distribuição dos dados, sendo bom para modelar relações não lineares complexas. Além disso, os classificadores podem ser sensíveis aos hiperparâmetros escolhidos, e como dito anteriormente, nesse trabalho foram usados os valores padrão para esses parâmetros.

O desempenho não tão satisfatório do kNN levantou algumas hipóteses: (i) As assinaturas podem simplesmente não servir para identificar entre proteínas não-nativas e nativas, apesar de conseguir classificar as funções das proteínas; (ii) O embaralhamento das cadeias pode não ser uma boa forma para geração das proteínas não-nativas; (iii) A quantidade de dados é muito pequena para um sistema tão complexo; (iv) O fato das assinaturas terem sido calculadas levando em consideração a estrutura inteira dos complexos pode ter gerado muito ruído nos dados.

Pretende-se investigar essas possibilidades em trabalhos futuros. Para a última, especificamente, pretende-se tentar outras formas de retratar a interação entre esses complexos, como usando a assinatura apenas da interface entre as proteínas, ou também concatenar a assinatura da interface com a assinatura de cada cadeia individualmente.

Referências

Basu, S. and Wallner, B. (2016a). DockQ: A quality measure for protein-protein docking models. *PLOS ONE*, 11(8):e0161879.

Basu, S. and Wallner, B. (2016b). Finding correct protein–protein docking models using ProQDock. *Bioinformatics*, 32(12):i262–i270.

Bertoline, L. M. F., Lima, A. N., Krieger, J. E., and Teixeira, S. K. (2023). Before and after AlphaFold2: An overview of protein structure prediction. *Frontiers in Bioinformatics*, 3.

Bryant, P., Pozzati, G., and Elofsson, A. (2022). Author correction: Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, 13(1).

Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., and Zupan, B. (2013). Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, 14:2349–2353.

Lensink, M. F., Méndez, R., and Wodak, S. J. (2007). Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins: Structure, Function, and Bioinformatics*, 69(4):704–718.

Moal, I. H., Torchala, M., Bates, P. A., and Fernández-Recio, J. (2013). The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC Bioinformatics*, 14(1).

Pires, D., Melo-Minardi, R., Santos, M., Da Silveira, C., Santoro, M., and Meira Jr, W. (2011a). Cutoff scanning matrix (csm): Structural classification and function prediction by protein inter-residue distance patterns. *BMC genomics*, 12 Suppl 4:S12.

Pires, D. E., de Melo-Minardi, R. C., dos Santos, M. A., da Silveira, C. H., Santoro, M. M., and Meira, W. (2011b). Cutoff scanning matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, 12(S4).

Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2013a). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3):335–342.

Pires, D. E. V., Blundell, T. L., and Ascher, D. B. (2015). pkCSM: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *Journal of Medicinal Chemistry*, 58(9):4066–4072.

Pires, D. E. V., de Melo-Minardi, R. C., da Silveira, C. H., Campos, F. F., and Meira, Wagner, J. (2013b). aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, 29(7):855–861.

Pires, D. E. V., de Melo-Minardi, R. C., da Silveira, C. H., Campos, F. F., and Meira, W. (2013c). aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, 29(7):855–861.

Rodrigues, C. H. M., Pires, D. E. V., and Ascher, D. B. (2021). pdCSM-PPI: Using graph-based signatures to identify protein–protein interaction inhibitors. *Journal of Chemical Information and Modeling*, 61(11):5438–5445.

Yin, R., Feng, B. Y., Varshney, A., and Pierce, B. G. (2022). Benchmarking scpAlphaFold/scp for protein complex modeling reveals accuracy determinants. *Protein Science*, 31(8).