Seleção de características utilizando Algoritmo Genético multiobjetivo e k-NN para predição de função de proteína

Bruno C. Santos, Cora Silberschneider, Marcos W. Rodrigues, Cristiane N. Nobre, Luis E. Zárate

Pontifícia Universidade Católica de Minas Gerais, Brasil
brunocs90@gmail.com, cora.silberschneider@sga.pucminas.br, marcoswanderrodrigues@gmail.com, {nobre,zarate}@pucminas.br

Abstract. The knowledge of a protein function is essential in many areas, such as bioinformatics, agriculture, and others. Therefore, it is necessary to provide efficient computational models that aim to find the function of a protein. Currently, there is a wealth of available information about protein, such as data from primary, secondary, tertiary and quaternary structures. One of the repositories that provide this information is the Sting DB, which has physicochemical information of the proteins, used by several authors. Our work proposes a methodology using the multiobjective genetic algorithm with non-parametric method k-NN during its genetic evolution, aiming to search the best subset of physical-chemical characteristics for the identification of protein classes. After that, we added new variables and applied PCA to the identified subset, to improve the classification process. In this step, we use the SVM due to its better performance with high dimensionalities data. The proposed methodology demonstrated accuracy values of 72.9% and an f-measure of 68.3%; also we gained about 90% efficiency in processing our approach compared to the previous model, allowing to add new attributes in an attempt to improve the prediction of protein function for future works.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Data Mining; I.2.6 [**Artificial Intelligence**]: Learning; J.3 [**Life and Medical Sciences**]: Biology and genetics

Keywords: Feature Selection, k-Nearest Neighbor, Multi-Objective Genetic Algorithm, Protein Prediction

1. INTRODUÇÃO

As proteínas são macromoléculas que existem abundantemente de formas variadas nas células. São formadas por cadeias polipeptídicas da combinação de aminoácidos e desempenham um papel fundamental no corpo humano, tendo funções construtoras e reparadoras do organismo, além de participar da formação dos hormônios, enzimas e anticorpos. Devido a esta importância, o conhecimento de sua função é fundamental para compreender os processos biológicos dos seres vivos.

Com o avanço das técnicas de sequenciamento genômico, o número de sequências de proteínas disponíveis para análise tem aumentado de forma significativa. No trabalho de Nadzirin and Firdaus-Raih [2012], os autores constataram que das proteínas que são descobertas, conhecemos a função de apenas 5% destas, cenário que persiste atualmente. Assim, é necessário o desenvolvimento de métodos computacionais para automatização e facilitação do processo de identificação da função proteica. Atualmente, existe uma quantidade considerável de métodos experimentais e computacionais para prever as funções de proteínas. No entanto, abordagens computacionais ainda não são capazes de predizer com precisão a função de uma extensa variedade de proteínas. Desse modo, o problema de predição de função da proteína permanece como um desafio para a biologia molecular e a bioinformática.

Uma proteína pode ser dividida em quatro níveis de acordo com sua estrutura, que são: 1) estrutura primária, composta por resíduos de aminoácidos unidos por ligações peptídicas; 2) estrutura

secundária que corresponde ao arranjo espacial de resíduos adjacentes em um segmento da cadeia polipeptídica; 3) estrutura terciária que ocorre quando resíduos distantes na cadeia polipeptídica se ligam após o enovelamento da proteína; e 4) estrutura quaternária a qual corresponde ao arranjo formado quando as proteínas possuem duas ou mais cadeias de aminoácidos [Lehninger et al. 2004]. Uma proteína também pode ser dividida de acordo com a função que desempenha, as funções são: regulatórias, transportadoras, contráteis e móveis, estruturais, protetoras e catalisadoras. Neste trabalho, optamos por utilizar as catalisadoras, a fim de comparar com outros trabalhos que também utilizaram esse mesmo tipo de proteínas. Essas desempenham tarefas de acelerar os processos biológicos e facilitar as reações químicas. As enzimas são um exemplo desta classe, correspondendo à maior classe de proteína, sendo conhecidas mais de 2000 tipos.

Este trabalho utiliza um conjunto de enzimas classificadas de forma hierárquica, em seis diferentes classes, de acordo com a reação química que catalisam. Essas enzimas recebem um identificador numérico chamado E.C *(Enzyme Commission)* criado pela IUBMB *(International Union of Biochemistry and Molecular Biology)*. O número E.C é composto por quatro níveis (1.2.3.4), sendo que o primeiro nível (1) informa qual das seis classes a enzima pertence: Oxidorredutases (E.C 1), Transferases (E.C 2), Hidrolases (E.C 3), Liases (E.C 4), Isomerases (E.C 5) e Ligases (E.C 6).

Inspirados nestes problemas da predição de função de proteína, apresentamos uma metodologia para seleção de características físico-químicas baseado em Algoritmo Genético (AG) multiobjetivo utilizando o classificador k-NN (*k-Nearest Neighbor*) durante a sua evolução genética e, posteriormente, enriquecendo o modelo com outras características. Por fim, utilizamos o método estatístico PCA *(Principal Component Analysis)* para a redução de dimensionalidade, em seguida a aplicação do classificador SVM (*Support Vector Machine*), e finalmente, a validação dos resultados encontrados.

O restante desse texto está organizado da seguinte forma: a Seção 2 traz os principais trabalhos relacionados à predição de proteínas. A Seção 3 descreve a metodologia, trazendo a descrição da base de dados, as etapas de pré-processamento e os métodos utilizados. A Seção 4 apresenta os resultados e discussões e, finalmente, a Seção 5 discute as conclusões finais do trabalho.

2. TRABALHOS RELACIONADOS

Em Yao and Ruzzo [2006] foi utilizado um *framework*, baseado no classificador k-NN, para predição da função de genes em dados heterogêneos. O autor afirma que o desempenho do k-NN está sujeito ao ajuste da métrica de similaridade. Para solucionar isso, o autor aplica métodos de regressão a fim de auxiliar na localização dos vizinhos mais prováveis de pertencerem à classe alvo. Além disso, o autor aplica o classificador SVM para integrar o dado heterogêneo, e afirma que ele possui melhor desempenho para encontrar genes que estão próximos dos limites de suas classes. Em comparação à SVM, o classificador k-NN possui vantagens de ter implementação mais simples, é computacionalmente eficiente, e produz informações adicionais que ajudam na interpretabilidade dos resultados.

Em Leijoto et al. [2014], os autores utilizaram um algoritmo genético mono objetivo para selecionar 11 características físico-químicas da base STING DB. Os valores de cada uma das variáveis (características físico-químicas) foram normalizados e aplicou-se a Transformada Discreta do Cosseno (TDC), considerando os 75 primeiros coeficientes. Para validar a abordagem, os autores utilizaram o classificador SVM com *Grid search* para ajustar os parâmetros *Cost* e γ do classificador. Foram realizados experimentos adicionando a frequência de cada aminoácido aos valores dos coeficientes da TDC, aumentando a sensibilidade e a precisão média do classificador para 68% e 71%, respectivamente. Como apontado pelos autores, o algoritmo genético teve limitação de processamento de 50 gerações e 10 indivíduos, devido à demanda do alto custo de processamento computacional.

Em Santos [2016], é feita uma avaliação das diferentes informações das quatro estruturas da proteína (primária, secundária, terciária e quaternária), os quais foram obtidas as seguintes informações: físico-químicas, potencial eletrostático, hidrofobicidade, frequência de aminoácidos, distâncias entre

carbonos α e peso molecular. Os valores das variáveis foram normalizados e a TDC considerou os 10 primeiros valores baseados em vários experimentos. Para a validação, o autor utilizou o classificador SVM com a abordagem *Grid search* ajustando os parâmetros *Cost* e γ. A metodologia proposta obteve valores médios de precisão de $78,4\%$ e sensibilidade de $74,3\%$. O autor comparou diferentes modelos baseados em SVM e conclui que todas as informações são relevantes para melhorar o desempenho do classificador. No entanto, o modelo considerou somente 10 de 344 características físico-químicas apontadas inicialmente por Mancini et al. [2004].

Em Santos et al. [2018] foi proposta uma metodologia utilizando algoritmo genético multiobjetivo para encontrar o subconjunto de características da base STING DB que melhor identifica as classes de enzimas. Após a seleção de atributos, realizou-se o enriquecimento da base com novas variáveis, de modo a construir um modelo baseado no classificador SVM. A metodologia proposta usando o AG obteve precisão de $77,3\%$ e *F-Measure* de $72,7\%$. Porém, toda a execução do AG utilizou o classificador SVM, o qual requer ajustes de parâmetros, tornando o seu processo oneroso computacionalmente. Com isso, os autores optaram por não realizar os ajustes de parâmetros durante a evolução do AG.

Neste trabalho, propomos uma metodologia utilizando o AG multiobjetivo para encontrar o subconjunto de características com o classificador k-NN, a fim de contornar o problema de ajustes de parâmetros do trabalho de Santos et al. [2018], e que melhor contribua para identificação das classes de enzimas estudadas. Após a seleção de atributos, adicionamos o enriquecimento com novas variáveis.

3. METODOLOGIA

Esta seção descreve a metodologia utilizada nesse trabalho. As etapas para a construção do modelo de predição de função de proteína baseado no classificador k-NN podem ser vistas na Figura 1.

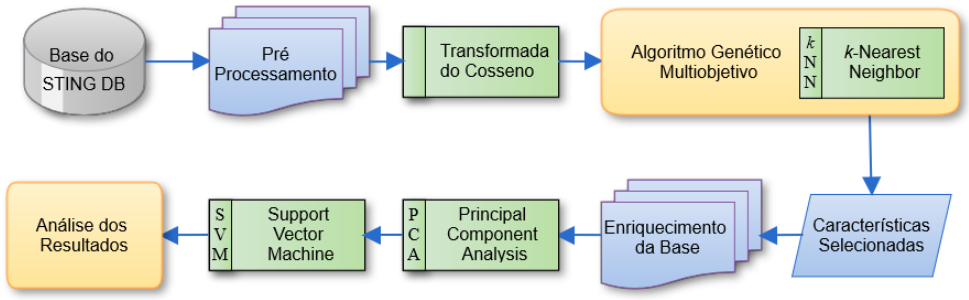

Fig. 1. Metodologia adotada.

3.1 Base de Dados

O conjunto de dados é formado pela base STING DB [Mancini et al. 2004], o qual permite extrair as principais características das seis enzimas investigadas neste trabalho, são elas: Hidrolase, Isomerase, Liase, Ligase, Oxidorredutase e Transferase. Tais enzimas foram usadas nos trabalhos de [Dobson and Doig 2004], [Borro et al. 2006], [Leijoto et al. 2014], [Santos 2016] e [Santos et al. 2018].

A base STING DB é um repositório desenvolvido pelo laboratório de Biologia Computacional da Embrapa Informática, o qual possui um conjunto de softwares e dados para visualizar e analisar a estrutura de macromoléculas. A base de dados possui um total de 490 cadeias de proteínas, distribuídas nas seis classes estudadas. Cada cadeia de proteína possui um total de 334 características provenientes do módulo *Java Protein Dossier* [Neshich et al. 2004]. Este módulo contém informações relativas às propriedades físico-químicas da proteína. Na Tabela I é possível observar o número de enzimas e

cadeias de proteínas utilizadas em nossa abordagem. Podemos observar que o número de proteínas é diferente do número de cadeias, pois uma mesma proteína pode ter duas ou mais cadeias polipeptídicas.

Tabela I. Classe e número de enzimas

	Proteínas usadas por Dobson and Doig		Proteínas após processo de limpeza	
Classe	Proteína	Cadeia	Proteína	Cadeia
Hidrolase	160	312	122	162
Isomerase	51	89	35	56
Liase	60	131	43	61
Ligase	20	22	15	16
Oxidorredutase	79	124	52	78
Transferase	128	162	82	117
Total	**498**	**840**	**349**	**490**

No trabalho de Santos [2016], os autores utilizaram essa mesma base de dados, o qual passou por um processo de limpeza, onde as enzimas com uma pontuação (*score*[1]) inferior a 0,3 foram eliminadas. As enzimas foram comparadas com as informações contidas no PDB (*Protein Data Bank*[2]) [Berman et al. 2000], o que permitiu observar que algumas destas enzimas foram classificadas em uma nova classe e, portanto, foram reorganizadas. Enzimas identificadas como obsoletas[3] no PDB não foram incluídas neste estudo. Isto resultou em uma redução no número de enzimas utilizadas nesta pesquisa, como é exibido nas últimas colunas da Tabela I.

3.2 Pré Processamento

Com o objetivo de aprimorar a qualidade das informações disponíveis, foi realizado o pré processamento dos dados, conforme a Figura 2. Inicialmente, foi realizada uma análise das características selecionadas, constatando-se a existência de dados redundantes, os quais foram removidos, restando um total de 291 características. Uma análise preliminar mostrou o alto custo computacional ao processar o conjunto de dados com 291 características, resultando em $\sum_{i=1}^{291} C_i^{291} = \frac{291!}{i!(291-i)!}$ combinações possíveis. Para reduzir a alta dimensionalidade do conjunto de dados, foi utilizada a técnica de correlação de *Pearson*. Com isso, notamos um grande número de características fortemente correlacionadas, e assim, optamos por eliminar as características que apresentaram correlação acima de 0.90, totalizando 69 características.

Fig. 2. Pré processamento dos dados

3.3 Transformada Discreta de Cosseno

Para que o uso de um classificador seja possível, o tamanho dos vetores de entrada deve ser o mesmo. No entanto, devido à diferença da quantidade de aminoácidos de cada cadeia de proteína, os vetores de entrada possuem tamanhos diferentes. Para contornar este problema, foi utilizada a técnica da Transformada Discreta do Cosseno (TDC) [Ahmed et al. 1974], o qual foi aplicada nas características

[1]https://scop.berkeley.edu/astral/spaci/ver=2.04
[2]http://www.rcsb.org/pdb/home/home.do
[3]Quando existe mudança nas coordenadas ou composição química de alguma proteína do PDB, esta é marcada como obsoleta e substituída por uma nova entrada e um novo identificador.

físico-químicas de todos os aminoácidos presentes na cadeia de aminoácidos que compõem a proteína. Foi escolhida a TDC por ela ser uma transformação que preserva nos valores iniciais dos coeficientes mais significativos, e nos restantes os valores que carregam pouca informação (Equação 1).

$$T_k = \alpha_k \sum^{N-1} X_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right], n > 0 \quad (1)$$

onde $\alpha_k = \frac{1}{\sqrt{N}} \forall k = 0$ ou $\alpha_k = \sqrt{\frac{2}{N}} \forall k = 1...N$, e N é número de aminoácidos de cada cadeia A_{ij}.

Com base em testes experimentais, foram definidos os $k = 10$ primeiros coeficientes da TDC por serem mais relevantes. Esse valor trouxe a melhor média para as medidas de precisão e sensibilidade, apontados em Santos [2016]. Com a aplicação da TDC todas as características físico-químicas das cadeias de proteínas possuem a mesma quantidade de registros de entrada, o que correspondente à $T_k = 10$ coeficientes. Logo, temos um conjunto final de dados com 490 cadeias de proteínas, onde cada uma delas possui um total de 69 atributos, representados cada um deles por 10 registros.

3.4 Algoritmo Genético Multiobjetivo

Após o processo de transformada dos dados, aplicamos o AG multiobjetivo NSGA-II (*Non-dominated Sorting Genetic Algorithm II*) a fim de buscar o melhor subconjunto de características com o menor percentual de erro do classificador, e usando o menor número de atributos para reduzir a complexidade do modelo. A sua escolha foi motivada por envolver diversos objetivos. A implementação foi na linguagem *Python* utilizando a biblioteca DEAP, disponível pela *Université Laval* [Fortin et al. 2012].

3.4.1 *Representação do indivíduo.* O indivíduo do AG representa uma possível solução para o problema a ser resolvido. Assim, cada indivíduo é representado por um vetor que possui 69 posições binárias, em que cada posição pode assumir valores entre 0 ou 1, indicando a presença ou ausência daquela determinada característica. Cada posição desse vetor representa uma característica extraída do STING DB, e para cada característica têm-se 10 coeficientes obtidos pela TDC.

3.4.2 *Função objetivo.* Dois aspectos foram considerados no processo de predição de função de proteína: 1) o modelo deve ter um percentual de erro baixo, de modo a aumentar a sua confiabilidade e, 2) o modelo deve ter um subconjunto pequeno de atributos para uma simplificação do modelo gerado. É importante notar que durante o processo de avaliação da função *fitness*, o critério de desempate de dois indivíduos foram: a) ter precisão mais alta, e b) o menor número de características. Para o cálculo da precisão utilizamos a técnica *Cross-Validation* com 10 partições. Estes critérios são detalhados a seguir:

—Menor percentual de erro médio de precisão do classificador k-NN, onde $Precisão = \frac{VP}{VP+FP}$, conforme Equação 2.

$$\overline{e_{Prec}} = 1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} Precisao_{ij}}{mn} \quad (2)$$

sendo $m = 6$ o número de classes de enzimas, e $n = 10$ o número de partições no *cross-validation*;

—Menor subconjunto de atributos dentre os 69 candidatos que melhor separa as classes de proteínas.

3.4.3 *Definição do tamanho da população.* De acordo com o trabalho de Santos et al. [2018], os melhores parâmetros encontrados durante a execução do AG multiobjetivo foram: População = 500, Gerações = 200, Cruzamento = $0,70\%$ e Mutação = $0,01\%$.

Com base nestes dados, realizamos vários experimentos para encontrar o subconjunto de características. Estes foram realizados com 5 sementes diferentes para cada parâmetro escolhido do AG, a fim

de garantir a confiabilidade dos resultados encontrados. Foram realizados 30 experimentos no total, de acordo com os intervalos de parâmetros estipulados e o número de sementes aleatórias, que são apresentados na Tabela II.

Tabela II. Parâmetros dos experimentos

Inicialização da população	:	Aleatória	Tamanho da população	:	500
Representação	:	Binária	Número de gerações	:	200, 300
Cruzamento	:	Dois pontos	Seleção Cruzamento	:	Torneio = 2
Cruzamento (Pc)	:	70%, 80%, 90%	Nova Geração	:	Não dominados
Mutação	:	Um ponto	Critério de parada	:	Número de gerações
Mutação (Pm)	:	1%			

3.5 Características Selecionadas

Após a execução dos experimentos, o AG encontrou um total de 26 características físico-químicas da base STING DB. Estas características podem ser vistas na Tabela III. A descrição detalhada destas características pode ser encontrada em Moraes et al. [2014].

Tabela III. Características selecionadas pelo AG

3DEntropyINT(6)	3DEntropyINT(9)	3DEntropyLHAsw(3,3)
ACCC()	Chi(0)	DensityCA(3)
DistanceC()	DistanceN()	EnergyDensityIFR(1)
EnergyDensityLHAsw(3,3)	EPsurface()	IFRSpongeCA(5)
InterfaceContactsEnergy(true,true)	InterfaceContactsEnergy(false,true)	InternalContactsEnergy(true,true)
NumberOfHBondPLC()	NumberofIFRContacts(2)	NumberofIFRContacts(3)
NumberofIFRContacts(9)	NumberofIFRContacts(11)	NumberofIFRContacts(13)
NumberofINTContacts(2)	NumberofINTContacts(3)	NumberofINTContacts(10)
NumberofINTContacts(12)	NumberofINTContacts(13)	

3.6 Enriquecimento e Aplicação do PCA e SVM

Os trabalhos de Santos [2016] e Santos et al. [2018] demonstraram que apenas as características físico-químicas da base STING DB não são suficientes para separar as 6 classes de enzimas. Portanto, utilizamos informações biológicas adicionais, visando melhorar os resultados. Estes atributos foram adicionados ao final da execução do AG, conforme descrito a seguir:

(1) Frequência de aminoácidos: para cada uma das cadeias consideradas, contabilizou-se a frequência com que cada um dos 20 aminoácidos aparece (20 características);
(2) Frequência do carbono *alpha*: padrão de distribuição da distância Euclidiana entre os carbonos α dos resíduos ao longo da cadeia (151 características) [Pires et al. 2011];
(3) Extração de dados estatísticos da estrutura primária: informações estatísticas sobre as sequências dos aminoácidos (31 características).

Com isso, obtivemos um total de 202 atributos adicionais para auxiliar na separação das seis classes de enzimas. Entretanto, observa-se que as características encontradas pelo AG, associadas às informações adicionais, possuem uma alta dimensionalidade: 462 atributos[4]. Com base nisto, aplicamos a Análise de Componentes Principais (PCA) para a reduzir o tamanho das entradas para o classificador, o que resultou em 117 componentes principais. Em seguida, aplicamos o classificador SVM. Vale ressaltar que, em testes preliminares, notou-se que o classificador k-NN, sobre o conjunto final, obteve

[4]Como o AG encontrou 26 características e cada uma das características é representada por 10 valores, temos: $26 * 10 = 260$. Acrescentando os atributos do enriquecimento, têm-se $260 + 202 = 462$ atributos.

uma baixa precisão. No entanto, o uso do k-NN durante o processamento do AG é recomendável devido ao seu desempenho computacional ser superior ao do SVM.

4. RESULTADOS

De acordo com o conjunto de características encontrados pelo AG com adição do enriquecimento da base, foi possível fixar os parâmetros $Cost = 4.0$ e $\gamma = 0.001953125$ fornecidos pelo algoritmo *Grid search* na execução da SVM, considerando as métricas de avaliação *Precisão*, *Sensibilidade* e *F-Measure* simples, como pode ser visto na Figura 3.

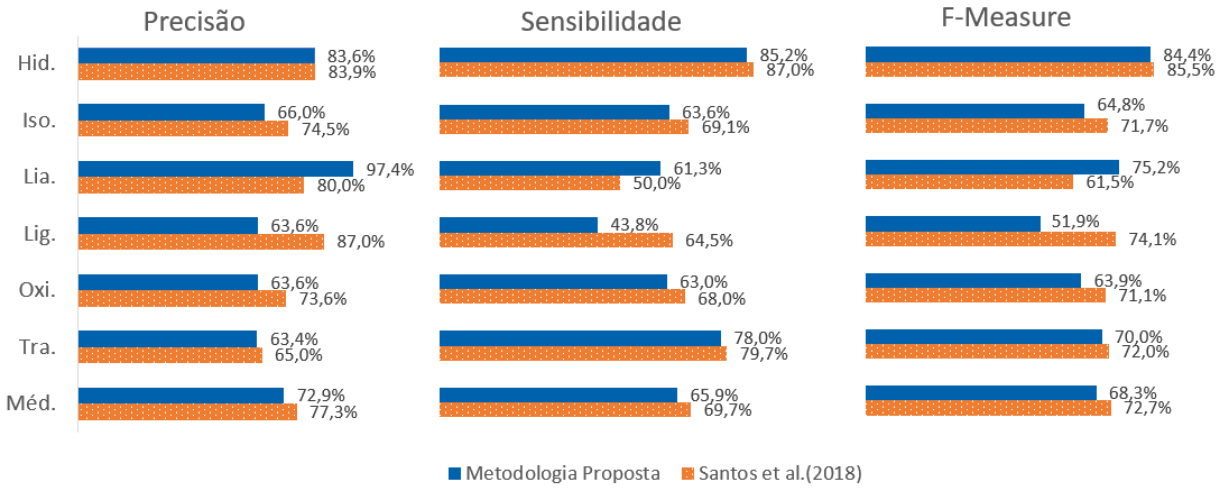

Fig. 3. Gráficos de Precisão, Sensibilidade e F-Measure

É possível observar uma variação de 4.4, 3.8 e 4.4 pontos percentuais menor nas médias de *Precisão*, *Sensibilidade* e *F-Measure* em relação ao trabalho de Santos et al. [2018]. No entanto, a redução nestas médias é devido ao fato de que o classificador k-NN não foi capaz de ajustar adequadamente o modelo, que manuseia uma base de dados de alta dimensão. Apesar disso, o tempo computacional utilizando k-NN diminuiu consideravelmente em relação ao uso do SVM. Com isso, houve um ganho e eficiência computacional no nosso modelo.

O uso do k-NN para seleção das melhores características, trouxe um benefício significante em relação ao tempo de execução dos testes. No trabalho de Santos et al. [2018] a otimização dos parâmetros do SVM demandaria um tempo computacional adicional de 30min, em média, para cada indivíduo, já com nossa abordagem não é necessário o ajuste de parâmetros. Com isso, cada geração do algoritmo genético demanda em média cerca de 1min[5], o que diminuiu em aproximadamente 90% o tempo de execução do modelo de predição em relação ao proposto por Santos et al. [2018].

5. CONCLUSÃO

Esse trabalho apresentou uma metodologia para a predição da função de proteína baseado no trabalho de Santos et al. [2018]. O modelo proposto utiliza o classificador supervisionado k-NN associado ao AG, a fim de selecionar os melhores atributos para a predição de função de proteínas. O objetivo principal foi obter melhores resultados, ou muito próximos dos resultados encontrados em Santos et al. [2018], porém, com um ganho significativo no tempo de processamento do modelo.

[5]Tempo gasto com uma População de 500 indivíduos. (Ex. P = 500, G = 200 x 1min = 200 min aproximadamente)

Apesar dos valores das médias serem parcialmente inferiores aos do trabalho anterior, é importante ressaltar o ganho acima de 90% na eficiência de processamento do modelo de predição com o uso do classificador k-NN. Este ganho é altamente significativo, uma vez que torna possível o enriquecimento da base pela adição de novas e/ou melhores características, a fim de melhorar a precisão da predição de proteínas.

Para trabalhos futuros, sugere-se realizar a seleção de características adicionando também dados do enriquecimento durante o processo evolucionário e a sua análise com métricas diferentes. Além disso, investigar os processos capazes de enriquecer a base de dados a fim de obter resultados mais precisos.

AGRADECIMENTOS

Os autores agradecem o apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) do Brasil.

REFERENCES

Ahmed, N., Natarajan, T., and Rao, K. R. Discrete cosine transform. *Computers, IEEE Transactions on* vol. C-23, pp. 90–93, 1974.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic Acids Research* vol. 28, pp. 235–242, 2000.

Borro, L. C., de Medeiros Oliveira, S. R., yamagishi, M. E. B., Mancini, A. L., Jardine, J. G., Mazoni, I., do Santos, E. H., Higa, R. H., Falcão, P. R. K., and Neshich, G. Predictiong enzyme class from protein structure using bayesian classification. *Genetic and Molecular Research* vol. 1, pp. 193–202, 2006.

Dobson, P. D. and Doig, A. J. Predicting enzyme class from protein structure without alignments. *Molecular Biology* vol. 345, pp. 187–199, 2004.

Fortin, F.-A., De Rainville, F.-M., Gardner, M.-A. G., Parizeau, M., and Gagné, C. Deap: Evolutionary algorithms made easy. *J. Mach. Learn. Res.* 13 (1): 2171–2175, July, 2012.

Lehninger, A., Nelson, D. L., and Cox, M. M. *Lehninger Principles of Biochemistry*. W. H. Freeman, 2004.

Leijoto, L., Assis De Oliveira Rodrigues, T., Zarate, L., and Nobre, C. A genetic algorithm for the selection of features used in the prediction of protein function. In *Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on*. pp. 168–174, 2014.

Mancini, A. L., Higa, R. H., Oliveira, A., Dominiquini, F., Kuser, P. R., Yamagishi, M. E. B., Togawa, R. C., and Neshich, G. Sting contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. *Bioinformatics* vol. 20, pp. 2145–2147, 2004.

Moraes, F. R., Neshich, I. A. P., Mazoni, I., Yano, I. H., Pereira, J. G. C., Salim, J. A., Jardine, J. G., and Neshich, G. Improving predictions of protein-protein interfaces by combining amino acid-specific classifiers based on structural and physicochemical descriptors with their weighted neighbor averages. *Plos One* 9 (1): 1–15, 2014.

Nadzirin, N. and Firdaus-Raih, M. Proteins of unknown function in the protein data bank (pdb): An inventory of true uncharacterized proteins and computational tools for their analysis. *International Journal of Molecular Sciences* 13 (10): 12761–12772, 2012.

Neshich, G., Rocchia, W., Mancini, A. L., Yamagishi, M. E. B., Kuser, P. R., Fileto, R., Baudet, C., Pinto, I. P., Montagner, A. J., Palandrani, J. F., Krauchenco, J. N., Torres, R. C., Souza, S., Togawa, R. C., and Higa, R. H. Javaprotein dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure. *Nucleic Acids Research* vol. 32, pp. W595–W601, 2004.

Pires, D. E., de Melo-Minardi, R. C., dos Santos, M. A., da Silveira, C. H., Santoro, M. M., and Meira, W. Cutoff scanning matrix (csm): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics* 12 (4): S12, 2011.

Santos, B. C., Nobre, C. N., and Zarate, L. E. Multi-objective genetic algorithm for feature selection in a protein function prediction context. In *IEEE Congress on Evolutionary Computation (CEC)*, 2018. (in press).

Santos, G. T. d. O. *Avaliação de características para predição de classes de enzimas com Support Vector Machine*. M.S. thesis, Pontifícia Universidade Católica de Minas Gerais, 2016.

Yao, Z. and Ruzzo, W. L. A regression-based k nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC Bioinformatics* 7 (1): S11, Mar, 2006.