doi 10.18699/bgrs2024-12.3-06

Поиск ДНК-связывающих белков (DBPs) с использованием методов глубокого обучения

Гавриленко А.[1]*, Шабурова Е.[1], Антонец Д.[1], Вяткин Ю.[1], Раменский В.[1, 2, 3]

[1] *Институт перспективных исследований проблем искусственного интеллекта и интеллектуальных систем, Московский государственный университет им. М.В. Ломоносова, Москва, Россия*

[2] *Национальный медицинский исследовательский центр терапии и профилактической медицины Минздрава России, Москва, Россия*

[3] *Факультет биоинженерии и биоинформатики, Московский государственный университет им. М.В. Ломоносова, Москва, Россия*

** a.gavrilenko@iai.msu.ru*

Ключевые слова: глубокое обучение; протеомика; языковые модели для белков; Ankh; ДНК-связывающие белки

Мотивация и цель: ДНК-связывающие белки (DBPs) играют ключевую роль во множестве биологических процессов, таких как репликация, репарация ДНК, регуляция транскрипции и трансляция [1]. Приблизительно 6–7 % генов в геномах эукариот и 2–5 % в геномах прокариот кодируют ДНК-связывающие белки [2]. Для изучения взаимодействия белков и ДНК используют несколько экспериментальных методов – поверхностный плазмонный резонанс, фильтрационный связывающий анализ, электрофоретический анализ сдвига подвижности [3]. Однако эти подходы имеют свои ограничения, являются длительными и трудоемкими, результаты экспериментов могут быть искажены. Достижения в области высокопроизводительных методов секвенирования значительно увеличили количество геномных данных. Однако менее 1 % из более чем 200 миллионов белков в UniProtKB имеют экспериментально подтвержденные аннотации. Это подчеркивает важность разработки вычислительных методов для автоматического поиска ДНК-связывающих белков (DBP). В последнее время в области вычислительной биологии получили распространение предобученные языковые модели для белков, многие из которых аналогичны по архитектуре языковым моделям для обработки естественного языка [4]. Они способны кодировать так называемые представления белка по его аминокислотной последовательности в виде вектора чисел, отражающего его свойства и функции. Эти представления используются для решения различных задач: классификация вторичной [5], третичной структуры белка [6], предсказания функции ферментов [7].

В настоящее время не существует общепринятых методов для предсказания способности белков связываться с ДНК. По этой причине использование предобученных белковых языковых моделей для решения данной задачи представляет значительный научный интерес. Целью нашей работы была разработка алгоритма для поиска белков, способных связываться с ДНК.

Методы и алгоритмы: Для формирования обучающего набора данных использовались аминокислотные последовательности белков из базы данных UniProtKB/Swiss-Prot [8]. Каждая запись в базе данных вручную проверяется и аннотируется специалистами, что обеспечивает высокую точность данных.

Задача поиска DBPs формулируется как задача бинарной классификации, где положительный класс составляют ДНК-связывающие белки, отрицательный класс – ДНК-несвязывающие белки.

Белки положительного класса выбирались из Swiss-Prot по имеющейся аннотации Gene Ontology “DNA binding”. Белки отрицательного класса также выбирались из Swiss-Prot, не имеющие аннотацию Gene Ontology “DNA-binding/RNA-binding”. РНК-связывающие белки не включались в отрицательный класс, поскольку многие из них также связываются с ДНК [9]. Для обеспечения уникальности набора данных были удалены идентичные последовательности белков. Белки фильтровались по длине аминокислотных последовательностей: исключались белки, имеющие длину менее 50 или более 1024 аминокислот. Этот подход был применен для удаления потенциальных пептидов и белковых комплексов, а также в связи с вычислительными ограничениями языковых моделей, которые не могут работать с длинными последовательностями белков [10]. Кроме того, исключались белки, имеющие неопределенные аминокислоты “X”. Тестовый набор данных PDB2272 был взят из предыдущего исследования [11], он представляют собой аминокислотные последовательности белков из UniprotKB с имеющейся аннотацией Gene Ontology. Данные использовались для оценки точности классификации разработанного алгоритма.

Алгоритм состоит из двух компонент:

1. Предобученная языковая модель для белков Ankh [12] для получения представлений белков по их аминокислотной последовательности.
2. Классификатор на основе градиентного бустинга, построенный с использованием LightAutoML [13], вычисляющий вероятность связывания ДНК белком.

Результаты: Для предотвращения гомологии между обучающим набором и тестовым набором данных была выполнена кластеризация с использованием MMseqs2 [14] с порогом идентичности аминокислотных последовательностей более чем 50 %. В результате белки из обучающего набора данных, попавшие в один кластер с белками из тестового набора, исключались. Для балансировки обучающего набора данных случайным образом выбирались белки отрицательного класса в эквивалентном соотношении к количеству белков положительного класса. Обучающий набор данных в итоге включил 31 803 белка положительного класса и 31 803 белка отрицательного класса. Информация о обучающем и тестовом наборах данных представлена в табл. 1.

Таблица 1. Статистика обучающего и тестового наборов данных

Набор данных	Количество белков			Количество кластеров
	общее	положительного класса	отрицательного класса	
Обучающий набор данных	63 606	31 803	31 803	28 033
PDB2272	2272	1119	1153	2270

Для оценки точности классификации использовались метрики: Precision, Sensitivity, Specificity, Matthews correlation coefficient (MCC). В ходе исследования

точность классификации алгоритма сравнивалась на тестовом наборе данных PDB2272 с существующими методами (табл. 2).

Таблица 2. Сравнение разработанного нами метода с существующими на тестовом наборе данных (PDB2272)

Алгоритм	Precision	Sensitivity	Specificity	MCC
PHMMER [15]	0.608	0.805	0.466	0.289
BiCaps-DBP [16]	*	0.893	0.707	0.613
Local-DPP [17]	*	0.087	0.936	0.456
Наш алгоритм	0.816	0.859	0.8	0.661

*Авторы работ не предоставили значения соответствующих метрик.

Выводы: В этом исследовании представлен новый метод использования предобученных языковых моделей для белков для идентификации ДНК-связывающих белков, расширяющий применение технологий машинного обучения в молекулярной биологии. Алгоритм показывает высокую эффективность по сравнению с существующими методами, значительно ускоряя и удешевляя процесс идентификации ДНК-связывающих белков. Это способствует более быстрому и точному пониманию их функций, с потенциальным применением в медицинских и биотехнологических исследованиях.

Searching for DNA-binding proteins (DBPs) using deep learning methods

Gavrilenko A.[1]*, Shaburova E.[1], Antonets D.[1], Vyatkin Y.[1], Ramensky V.[1, 2, 3]

[1] *Institute for Artificial Intelligence, Lomonosov Moscow State University, Moscow, Russia*

[2] *National Medical Research Center for Therapy and Preventive Medicine of the Ministry of Healthcare of Russian Federation, Moscow, Russia*

[3] *Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia*

** a.gavrilenko@iai.msu.ru*

Key words: deep learning; proteomics; language models for proteins; Ankh; DNA-binding proteins

Motivation and Aim: DNA-binding proteins (DBPs) play a crucial role in numerous biological processes such as DNA replication, repair, transcriptional regulation, and translation [1]. Approximately 6–7 % of genes in eukaryotic genomes and 2–5 % in prokaryotic genomes encode DNA-binding proteins (DBPs) [2]. Several experimental methods are employed to study protein-DNA interactions, including surface plasmon resonance, filter binding assays, and electrophoretic mobility shift assays [3]. However, these approaches have limitations, are time-consuming and labor-intensive, and experimental results may be biased. Advances in high-throughput sequencing methods have significantly increased the amount of genomic data available. Nevertheless, less

than 1 % of the over 200 million proteins in UniProtKB have experimentally validated annotations. This underscores the importance of developing computational methods for the automatic identification of DNA-binding proteins (DBPs). Recently, pretrained protein language models have become popular in computational biology, many of which are analogous in architecture to natural language processing models [4]. These models can encode protein representations from their amino acid sequences into vector forms reflecting their properties and functions. These representations are used for various tasks, such as the classification of secondary [5] and tertiary protein structures [6], and the prediction of enzyme functions [7].

Currently, no standardized methods exist for predicting the DNA-binding capability of proteins. For this reason, utilizing pretrained protein language models for this task presents significant scientific interest. The aim of this study was to develop an algorithm for identifying proteins capable of binding DNA.

Methods and Algorithms: Amino acid sequences of proteins from the UniProtKB/Swiss-Prot database [8] were used to form the training dataset. Each entry in the database is manually curated and annotated by experts, ensuring high data accuracy. The task of identifying DBPs is formulated as a binary classification problem, where the positive class comprises DNA-binding proteins and the negative class comprises non-DNA-binding proteins.

Proteins in the positive class were selected from Swiss-Prot based on the existing Gene Ontology annotation "DNA binding." Proteins in the negative class were also selected from Swiss-Prot, excluding those with the Gene Ontology annotation "DNA-binding/RNA-binding." RNA-binding proteins were not included in the negative class, as many of them also bind to DNA [9]. To ensure dataset uniqueness, identical protein sequences were removed. Proteins were filtered by the length of their amino acid sequences, excluding those shorter than 50 or longer than 1,024 amino acids. This approach was applied to remove potential peptides and protein complexes and due to computational limitations of language models, which cannot process long protein sequences [10]. Additionally, proteins with undefined amino acids "X" were excluded. The test dataset PDB2272 was taken from a previous study [11] and consists of amino acid sequences of proteins from UniProtKB with existing Gene Ontology annotations. The data were used to assess the classification accuracy of the developed algorithm.

The algorithm comprises two components:

1. A pretrained protein language model, Ankh [12], to obtain protein representations from their amino acid sequences.
2. A gradient boosting-based classifier built using LightAutoML [13], which calculates the probability of a protein binding to DNA.

The algorithm consists of two components:

1. A pre-trained protein language model, Ankh [12], to extract features from the amino acid sequences.
2. A gradient boosting classifier, built using LightAutoML [13], to predict the probability of a protein being DNA-binding.

Results: To avoid homology between the training and test datasets, clustering was performed using MMseqs2 [14] with an amino acid sequence identity threshold of over 50 %. As a result, proteins from the training dataset that clustered with proteins from the test dataset were excluded. To balance the training dataset, proteins from the negative class were randomly selected in a ratio equivalent to the number of proteins in the positive class. The final training dataset included 31,803 proteins in the positive class

and 31,803 proteins in the negative class. Information about the training and test datasets is presented in Table 1.

Table 1. Statistics of the training and test datasets

Dataset	Number of Proteins	Number of Positive Class Proteins	Number of Negative Class Proteins	Number of Clusters
Training dataset	63,606	31,803	31,803	28,033
PDB2272	2,272	1,119	1,153	2,270

Metrics such as Precision, Sensitivity, Specificity, and Matthews correlation coefficient (MCC) were used to evaluate classification accuracy. During the study, the classification accuracy of the algorithm was compared with existing methods on the PDB2272 test dataset (Table 2).

Table 2. Comparison with existing methods on an independent test set (PDB2272)

Algorithm	Precision	Sensitivity	Specificity	MCC
PHMMER [15]	0.608	0.805	0.466	0.289
BiCaps-DBP [16]	*	0.893	0.707	0.613
Local-DPP [17]	*	0.087	0.936	0.045
Our Algorithm	0.816	0.859	0.800	0.661

*The authors did not provide values for corresponding metrics.

Conclusion: This study introduces a novel method using pretrained protein language models to identify DNA-binding proteins, extending machine learning applications in molecular biology. The algorithm shows high efficiency compared to existing methods, significantly accelerating and reducing the cost of identifying DNA-binding proteins. This contributes to a faster and more accurate understanding of their functions, with potential applications in medical and biotechnological research.

Список литературы/References

1. Ahmed S., Kabir M., Ali Z. et al. An Integrated Feature Selection Algorithm for Cancer Classification using Gene Expression Data. *Comb Chem High Throughput Screen.* 2018;21(9):631-645
2. Luscombe N.M., Austin S.E., Berman H.M., Thornton J.M. An overview of the structures of protein-DNA complexes. *Genome Biol*. 2000;1(1):REVIEWS001
3. Ferraz R.A.C., Lopes A.L.G., da Silva J.A.F. et al. DNA-protein interaction studies: A historical and comparative analysis. *Plant Methods.* 2021;17(1):82
4. Ofer D., Brandes N., Linial M. The language of proteins: NLP, machine learning & protein sequences. *Comput Struct Biotechnol J.* 2021;19:1750-1758

5. Rives A., Meier J., Sercu T. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA.* 2021;118(15):e2016239118
6. Chowdhury R, Bouatta N, Biswas S. et al. Single-sequence protein structure prediction using a language model and deep learning. *Nat Biotechnol.* 2022;40(11):1617-1623
7. Yu T., Cui H., Li J.C. et al. Enzyme function prediction using contrastive learning. *Science*. 2023;379(6639):1358-1363
8. UniProt Consortium T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2018;46(5):2699
9. Hudson W.H., Ortlund E.A. The structure, function and evolution of proteins that bind DNA and RNA. *Nat Rev Mol Cell Biol.* 2014;15(11):749-760
10. Tan N.Ö., Peng A.Y., Bensemann J. et al. Input-length-shortening and text generation via attention values. *arXiv*. 2023
11. Lou W., Wang X., Chen F. et al. Sequence based prediction of dna-binding proteins based on hybrid feature selection using random forest and gaussian naïve bayes. *PLoS One.* 2014;9(1):e86703
12. Elnaggar A., Essam H., Salah-Eldin W. et al. Ankh: Optimized protein language model unlocks general-purpose modelling. *arXiv*. 2023
13. Vakhrushev A., Ryzhkov A., Savchenko M. et al. LightAutoML: AutoML solution for a large financial services ecosystem. *arXiv*. 2021
14. Steinegger M., Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol.* 2017;35(11):1026-1028
15. Potter S.C., Luciani A., Eddy S.R. et al. HMMER web server: 2018 update. *Nucleic Acids Res.* 2018;46(W1):W200-W204
16. Mursalim M.K.N., Mengko T.L.E.R., Hertadi R. et al. BiCaps-DBP: Predicting DNA-binding proteins from protein sequences using Bi-LSTM and a 1D-capsule network. *Comput Biol Med.* 2023;163:107241
17. Wei L., Tang J., Zou Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. *Inf Sci.* 2017;384:135-144