doi 10.18699/bgrs2024-8.4-02

Предсказание количества белка в дрожжах, основанное на трансферном обучении

Вензель А.С.[1, 2]*, Клименко А.И.[1, 2], Деменков П.С.[1, 2, 3], Иванисенко Т.В.[1, 2, 3], Лашин С.А.[1, 2, 3], Колчанов Н.А.[1, 2, 3], Иванисенко В.А.[1, 2, 3]

[1] *Институт цитологии и генетики СО РАН, Новосибирск, Россия*

[2] *Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия*

[3] *Новосибирский государственный университет, Новосибирск, Россия*

** venzel@bionet.nsc.ru*

Ключевые слова: дрожжи; количество белка; трансформеры; машинное обучение

Прогнозирования уровня белков в бактериях имеет большое значение для биотехнологии и оптимизации процессов биосинтеза, в частности для компьютерного дизайна штаммов-продуцентов с повышенной экспрессией целевых наборов белков. Использование современных подходов машинного обучения улучшает понимание количественной динамики белков и ускоряет разработку новых биотехнологических приложений. И существующие современные методы глубокого обучения, такие как нейросетевые модели обработки естественного языка с архитектурой Transformer, обученные на больших массивах данных, позволяют анализировать биологические последовательности без «ручного» извлечения признаков последовательности, что значительно повышает точность предсказаний моделей машинного обучения. В рамках данной работы был разработан метод предсказаний количества белка в *Saccharomyces cerevisiae* S288C, названный yeastProtPred, включающий в себя совместное использованеи аминокислотных и геномных трансформеров для получения векторных представлений последовательностей гена, а также последовательностей кодируемого белка.

Информация о полном геноме дрожжей Saccharomyces cerevisae S288C и аннотация этого генома были взяты из базы данных The Saccharomyces Genome Database (SGD). Количественные данные по уровням представленности белков в клетках *Saccharomyces cerevisiae* S288C были взяты из Protein Abundance Database (PaxDb). Для получения векторных представлений последовательностей генов и белков использовались предобученные BERT-образные трансформеры GENA-LM и ESM-2 соответственно. Векторные представления используются, как входные параметры для обучения модели градиентного бустинга для предсказания количества белка.

Финансирование: Исследование проведено при финансовой поддержке проекта Министерства науки и высшего образования РФ «Курчатовский центр геномных исследований мирового уровня» № 075-15-2019-1662 от 31.10.2019.

Transfer learning based yeast protein abundance prediction

Venzel A.S.[1, 2]*, Klimenko A.I.[1, 2], Demenkov P.S.[1, 2, 3], Ivanisenko T.V.[1, 2], Lashin S.A.[1, 2, 3], Kolchanov N.A.[1, 2, 3], Ivanisenko V.A.[1, 2, 3]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[2] *Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[3] *Novosibirsk State University, Novosibirsk, Russia*

* *venzel@bionet.nsc.ru*

Key words: yeast; protein abundance; transformer; machine learning

Protein abundance prediction in bacteria is crucial for biotechnology and optimizing biosynthesis processes, particularly for computer-aided producing strains design with enhanced target gene expressions. Modern machine learning approaches enhance our understanding of protein quantitative dynamics and accelerate the development of new biotechnological applications. Existing deep learning methods, such as natural language processing neural network models with Transformer architecture trained on large datasets, allow to analyze biological sequences without manual feature extraction, significantly improving the accuracy of machine learning model predictions.

In this study, a method named yeastProtPred was developed for predicting protein abundance in Saccharomyces cerevisiae S288C. It incorporates the joint use of amino acid and genomic transformers to obtain vector representations of gene sequences and encoded protein sequences.

Information about the complete genome of Saccharomyces cerevisiae S288C and its annotation were obtained from The Saccharomyces Genome Database (SGD). Quantitative data on protein abundance levels in Saccharomyces cerevisiae S288C cells were taken from the Protein Abundance Database (PaxDb). Pre-trained BERT-like transformers GENA-LM and ESM-2 were used to obtain embeddings of gene and protein sequences, respectively. These embeddings are used as input vectors for training a gradient boosting model to predict protein abundance.

Funding: The study was funded by the Ministry of Science and Higher Education of the Russian Federation project “Kurchatov Center for World-Class Genomic Research” No. 075-15-2019-1662 from 2019-10-31.