Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1145-1154

doi 10.18699/vjgb-25-119

# OrthoML2GO: предсказание функций белков по гомологии с использованием ортогрупп и алгоритмов машинного обучения

Е.В. Малюгин[1], Д.А. Афонников[2]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
evgeny.malyugin98@gmail.com

**Аннотация.** В последние годы быстрый рост объемов данных секвенирования обострил проблему функциональной аннотации белковых последовательностей, поскольку традиционные методы, основанные на гомологии, сталкиваются с ограничениями при работе с отдаленными гомологами, что затрудняет наиболее точное определение функций белков. В нашей работе представлен метод предсказания функций белков OrthoML2GO, который интегрирует поиск гомологичных последовательностей с помощью алгоритма USEARCH, анализ ортогрупп на базе OrthoDB 12-й версии и алгоритм машинного обучения (градиентный бустинг). Ключевая особенность подхода заключается в использовании информации об ортогруппах для учета эволюционного и функционального сходства белков и применения машинного обучения для дальнейшего уточнения терминов Gene Ontology (GO) для анализируемой последовательности. Для выбора оптимального алгоритма аннотации белков были поэтапно применены следующие подходы: метод *k* ближайших соседей (KNN); метод на основе аннотации ортогруппы, наиболее представленной у *k* ближайших гомологов (OG); метод верификации выявленных на предыдущем этапе терминов GO с помощью алгоритмов машинного обучения. Проведено сравнение точности предсказания терминов GO методом OrthoML2GO с программами аннотации Blast2GO и PANNZER2 на выборках последовательностей как отдельных организмов (человек, арабидопсис), так и на комбинированной выборке последовательностей, представленных разными таксонами. Результаты показали, что предложенный метод не уступает, а по некоторым показателям превосходит их по качеству предсказания функций белков, особенно на больших и разнородных выборках организмов, а наибольший прирост точности достигается за счет комбинации информации о ближайших гомологах и ортогруппах в сочетании с верификацией терминов методами машинного обучения. Разработанный подход демонстрирует высокую эффективность для крупномасштабной автоматической аннотации белков. Перспективы дальнейшего развития включают оптимизацию параметров моделей машинного обучения под конкретные биологические задачи и интеграцию дополнительных источников структурно-функциональной информации, что позволит еще больше повысить точность и универсальность метода. Кроме того, внедрение новых инструментов биоинформатики и расширение базы данных аннотированных белков будут способствовать дальнейшему совершенствованию предложенного подхода.

**Ключевые слова:** предсказание функций белка; генная онтология; гомология; ортогруппа; машинное обучение

**Для цитирования:** Малюгин Е.В., Афонников Д.А. OrthoML2GO: предсказание функций белков по гомологии с использованием ортогрупп и алгоритмов машинного обучения. *Вавиловский журнал генетики и селекции*. 2025;29(7):1145-1154. doi 10.18699/vjgb-25-119

**Финансирование.** Исследование поддержано Курчатовским геномным центром ИЦиГ СО РАН, соглашение с Министерством образования и науки Российской Федерации № 075-15-2019-1662 и бюджетным проектом № FWNR-2022-0020.

**Благодарности.** Исследование выполнено с использованием суперкомпьютерного комплекса ЦКП «Биоинформатика» ИЦиГ СО РАН.

# OrthoML2GO: homology-based protein function prediction using orthogroups and machine learning

E.V. Malyugin[1], D.A. Afonnikov[2]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
evgeny.malyugin98@gmail.com

**Abstract.** In recent years, the rapid growth of sequencing data has exacerbated the problem of functional annotation of protein sequences, as traditional homology-based methods face limitations when working with distant homologs, making it difficult to accurately determine protein functions. This paper introduces the OrthoML2GO method for

protein function prediction, which integrates homology searches using the USEARCH algorithm, orthogroup analysis based on OrthoDB version 12.0, and a machine learning algorithm (gradient boosting). A key feature of our approach is the use of orthogroup information to account for the evolutionary and functional similarity of proteins and the application of machine learning to refine the assigned GO terms for the target sequence. To select the optimal algorithm for protein annotation, the following approaches were applied sequentially: the k-nearest neighbors (KNN) method; a method based on the annotation of the orthogroup most represented in the k-nearest homologs (OG); a method of verifying the GO terms identified in the previous stage using machine learning algorithms. A comparison of the prediction accuracy of GO terms using the OrthoML2GO method with the Blast2GO and PANNZER2 annotation programs was performed on sequence samples from both individual organisms (humans, Arabidopsis) and a combined sample represented by different taxa. Our results demonstrate that the proposed method is comparable to, and by some evaluation metrics outperforms, these existing methods in terms of the quality of protein function prediction, especially on large and heterogeneous samples of organisms. The greatest performance improvement is achieved by combining information about the closest homologs and orthogroups with verification of terms using machine learning methods. Our approach demonstrates high performance for large-scale automatic protein annotation, and prospects for further development include optimizing machine learning model parameters for specific biological tasks and integrating additional sources of structural and functional information, which will further improve the method's accuracy and versatility. In addition, the introduction of new bioinformatics tools and the expansion of the annotated protein database will contribute to the further improvement of the proposed approach.
**Key words:** protein function prediction; gene ontology; homology; orthogroup; machine learning

**For citation:** Malyugin E.V., Afonnikov D.A. OrthoML2GO: homology-based protein function prediction using orthogroups and machine learning. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1145-1154. doi 10.18699/vjgb-25-119

## Введение

Внедрение технологий секвенирования нового поколения (NGS) привело к экспоненциальному росту объемов данных о последовательностях ДНК, РНК и белков (Goodwin et al., 2016). Основными источниками этих данных служат масштабные и многочисленные проекты в области геномики, транскриптомики и протеомики (Cheng et al., 2018; Lewin et al., 2018). Однако значительная доля идентифицируемых в рамках таких проектов последовательностей остается с неизвестной функцией (Galperin, Koonin, 2010).

Экспертная аннотация генов требует большого количества времени для поиска информации о функциях генов в литературных источниках и, хотя является наиболее достоверной, использовать ее для огромного количества вновь предсказанных генов невозможно. Поэтому для большинства новых аминокислотных последовательностей (далее для краткости – последовательности) необходима разработка эффективных методов автоматического аннотирования, позволяющих определять их молекулярные функции, роль в клеточных процессах и клеточную локализацию. С учетом широкого использования для аннотации функций базы данных онтологии генов (Gene Ontology, GO) (Ashburner et al., 2000; Du Plessis et al., 2011; Gene Ontology Consortium, 2023) задача сводится к присвоению последовательности этих терминов в автоматическом режиме.

Большинство методов прогнозирования функций белков, основанных на анализе последовательности или трехмерной структуры, опирается на фундаментальный принцип: функция может быть предсказана на основе установления достоверного структурного или эволюционного сходства с белком, функция которого уже известна (Benso et al., 2013). Важнейшей задачей здесь является расшифровка взаимосвязи между обнаруженным сходством структур или последовательностей и фактическим уровнем функционального родства (Pearson, 2013). Среди этих методов к числу наиболее популярных относятся методы предсказания функции по гомологии, благодаря их универсальности и относительной простоте. Методы, основанные на гомологии, присваивают анализируемому белку термины GO на основе сходства его аминокислотной последовательности с первичными структурами белков с известной функцией. Другими словами, функция белка может быть расшифрована путем анализа его сходства с другими белками, для которых надежно определена функция (Eisenberg et al., 2000; Pearson, 2013).

Для сравнения аминокислотных последовательностей двух белков широко используется метод BLAST (Altschul et al., 1990). Однако в последнее время стали появляться новые инструменты поиска гомологичных последовательностей в базах данных, такие как GHOSTX (Suzuki et al., 2014), DIAMOND (Buchfink et al., 2015), MMseqs2 (Steinegger, Söding, 2017) и многие другие. Их характерная особенность – скорость обработки информации, на порядки превышающая скорость работы программ пакета BLAST, достигаемая в большинстве своем за счет более эффективной обработки совпавших фрагментов последовательности.

Концепция гомологии является основополагающей для получения выводов о процессах эволюционного формирования генов. В 1970 г. Уолтер Фитч (Fitch, 1970) предложил классифицировать гомологичные белки на ортологи и паралоги в соответствии с их происхождением. Ортологи происходят в процессе эволюционной дивергенцией генов, относящихся к различным таксонам в ходе их видообразования. Паралоги образуются за счет дупликаций генов. Предполагается, что ортологи сохраняют функцию гена-предшественника предкового вида, в то время как паралогичные гены после дупликаций могут изменить функцию (Fitch, 2000; Kuzniar et al., 2008; Altenhoff et al., 2019). С учетом огромной важности ортологов для сравнительной геномики и функциональной аннотации информация об ортологичных генах и их семействах накапливается в ряде специализированных баз данных, которые предоставляют собой важнейшие источники информации для идентификации и анализа ортологичных

**Таблица 1.** Список организмов, вошедших в исследование

| Организм | Число последовательностей | Источник аннотации |
|---|---|---|
| *Arabidopsis thaliana* | 27 655 | TAIR (Reiser et al., 2024) |
| *Homo sapiens* | 19 763 | EBI Gene Ontology Annotation Database (Huntley et al., 2015) |
| *Drosophila melanogaster* | 28 543 (с изоформами) | FlyBase (Öztürk-Çolak et al., 2024) |
| *Solanum tuberosum* | 40 722 (с изоформами) | SpudDB (Hamilton et al., 2025a) |
| *Danio rerio* | 33 428 (с изоформами) | ZFIN (Bradford et al., 2022) |
| *Chlamydomonas reinhardtii* | 16 090 | PhycoCosm (Grigoriev et al., 2021) |
| *Oryza sativa* | 34 226 (с изоформами) | RGAP (Hamilton et al., 2025b) |

групп генов (ортогрупп) (Jensen et al., 2008; Kriventseva et al., 2008). Отметим, что для решения задач предсказания функций генов успешно используются методы, включающие алгоритмы машинного обучения, которые позволяют увеличить точность по сравнению с более ранними подходами (Sanderson et al., 2023; Yuan et al., 2023).

В настоящей работе исследуется возможность предсказания функций белков на основе поиска гомологичных последовательностей, учета их ортологов и методов машинного обучения. Проведен поэтапный анализ влияния трех указанных факторов на точность предсказания терминов GO. Показано, что среди методов машинного обучения наибольшую точность предсказания демонстрирует алгоритм градиентного бустинга. На этой основе реализован алгоритм предсказания OrthoML2GO. Его точность сравнили с методами Blast2GO и PANNZER2. Показано, что предложенный метод обеспечивает более высокую точность, особенно на больших и разнородных выборках.

## Материалы и методы

**Данные об аминокислотных последовательностях.** Перечень видов организмов и аминокислотных последовательностей, использованных в работе, дан в табл. 1. Он включает организмы с разной степенью полноты аннотации геномов (табл. S1[1]), представляющих разные таксоны как растений, так и животных: двудольных, однодольных, одноклеточных водорослей, позвоночных, членистоногих (см. табл. 1).

**База данных OrthoDB как источник гомологичных последовательностей, аннотации и информации об ортологии.** Мы использовали базу данных OrthoDB v 12.0 (https://www.orthodb.org/) (Tegenfeldt et al., 2025) в качестве источника гомологичных последовательностей, их аннотации терминами GO и данных по ортологии. База данных включает информацию по 5827 видам эукариот, 17 551 бактерий, 607 архей и 7962 вирусов. Она содержит более 162 млн последовательностей, классифицированных более чем на 10 млн групп ортологов. База данных также включает GO-аннотацию для части последовательностей и представляет собой удобный источник их классификации на ортологи и GO-аннотации. Кроме того, эта база данных предоставляет классификацию белковых последовательностей на ортологические семейства, для которых также представлена обобщенная аннотация функций белков в терминах GO.

**Поиск гомологичных последовательностей.** Поиск гомологов проводился с помощью алгоритма USEARCH v 11.0.667 (https://drive5.com/usearch/) (Edgar, 2010) командой usearch_local. Она производит поиск высокоидентичных совпадений на несколько порядков быстрее BLAST. В процессе поиска гомологичных последовательностей неизбежно оказывалось, что список гомологов включал и саму искомую последовательность. Для объективной оценки мы исключали из результатов поиска такие идентичные последовательности.

**Общая схема аннотации последовательностей.** Поиск терминов GO был реализован на языках bash Linux и R на вычислительных ресурсах центра коллективного пользования «Биоинформатика» ИЦиГ СО РАН. Было разработано три алгоритма аннотации функций белков на основе базы данных OrthoDB (рис. 1).

Слева (см. рис. 1, *а*) в большом овале схематически показана база OrthoDB v 12.0 (Tegenfeldt et al., 2025) с представителями ортологических групп (ортогрупп) OG1…OG3. Последовательности ортологичных семейств показаны на рисунке прямоугольниками одного цвета. Первый, базовый алгоритм предсказания последовательностей основан на поиске $k$ ближайших гомологов и обозначается KNN. С помощью программы USEARCH в базе OrthoDB для анализируемой последовательности ведется поиск гомологичных последовательностей и ранжируются по уровню сходства. Они могут включать как представителей одной ортогруппы, так и других (показаны разными цветами). Анализируемой последовательности присваиваются термины GO $k$ наиболее сходных последовательностей (см. рис. 1, *б*).

Второй метод основан на принципе ортологии и будет обозначен как OG. Для каждого из $k$ ближайших гомологов анализируемой последовательности определяется его ортогруппа в БД OrthoDB. Ортогруппа, к которой принадлежит анализируемая последовательность, определяется методом голосования: это ортогруппа, имеющая наибольшую частоту встречаемости среди всех $k$ ближайших гомологов (см. рис. 1, *в*). Термины GO для последовательностей из этой ортогруппы присваиваются анализируемой последовательности (см. рис. 1, *г*).

Третий подход, обозначенный как KNN+OG (см. рис. 1, *д*), заключается в объединении терминов GO, полученных с помощью предыдущих алгоритмов – KNN

[1] Табл. S1–S12 Приложения см. по адресу:
https://vavilovj-icg.ru/download/pict-2025-29/appx43.pdf

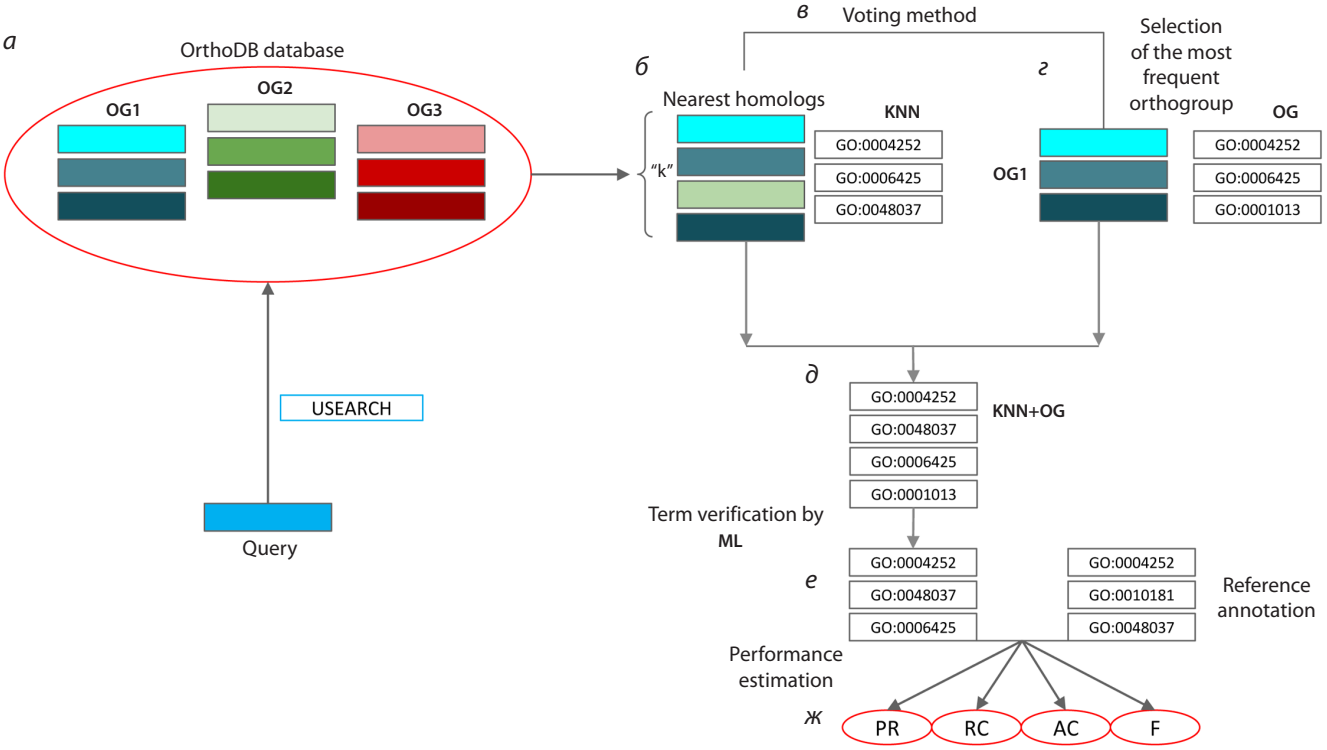


**Рис. 1.** Общая схема аннотации последовательностей и ее оценки. Последовательности, принадлежащие одной ортогруппе, представлены разными оттенками одного цвета: синего, зеленого или красного.

*а* – база данных OrthoDB с ортогруппами; *б* – присвоение терминов GO от *k* ближайших гомологов (метод KNN); *в* – определение наиболее представленной ортогруппы методом голосования; *г* – присвоение терминов GO, ассоциированных с выбранной ортогруппой (метод OG); *д* – объединение терминов GO, полученных методами KNN и OG (метод KNN+OG); *е* – верификация объединенного списка терминов с помощью машинного обучения; *ж* – сравнение предсказанных терминов с референсной аннотацией и расчет метрик.

и OG – для анализируемой последовательности (см. рис. 1, *е*). Этот список терминов GO, который сравнивается с референсной аннотацией (истинной) при помощи таких мер, как: precision, recall (sensitivity), accuracy и F-score (F-measure), которая была результирующей (см. рис. 1, *ж* и раздел «Верификация терминов методами машинного обучения»).

**Методы аннотации анализируемой последовательности терминами GO.** ***Метод k ближайших гомологов*** **(KNN).** *k* ближайших гомологов по уровню сходства определяются в результате поиска по БД OrthoDB программой USEARCH с параметрами: identity (идентичность аминокислотных последовательностей) = 50 %, coverage (покрытие анализируемой последовательности найденным гомологом) = 70 %, *e*-value (статистическая значимость найденного совпадения) = $10^{-6}$, что оправдано целью уменьшить ложноположительные находки на этапе поиска гомологов. Анализируемой последовательности присваиваются термины GO *k* наиболее сходных последовательностей, имеющихся в БД OrthoDB. Значения параметра *k* могут варьировать (Kharsikar et al., 2007; Dongardive, Abraham, 2016), поэтому в нашем случае из интервала *k* = 1–30 с шагом 5 мы выбирали значение, при котором точность идентификации терминов с использованием аннотации OrthoDB (табл. S4–S9) была наилучшей.

***Использование ортологических групп*** **(OG).** В этом методе для каждого из *k* ближайших гомологов анализируемой последовательности, определенного методом KNN, согласно аннотации OrthoDB, выбирается ортологическая группа, соответствующая наиболее древнему предковому таксону. После этого определяется наиболее часто встречающаяся среди *k* ортогруппа, которая присваивается анализируемой последовательности. Термины аннотации GO для последовательностей из этой ортогруппы в БД OrthoDB присваиваются анализируемой последовательности. Метод KNN+OG основан на объединении терминов GO (с исключением повторов), полученных отдельно методами KNN и OG, описанными выше.

***Верификация терминов методами машинного обучения.*** Для уточнения списка предсказанных терминов GO на третьем этапе анализа (см. рис. 1, *е*) использовали три алгоритма машинного обучения (ML – machine learning): логистическую регрессию (LR), метод градиентного бустинга (XGB) и метод случайного леса (RF). Отметим, что этот этап не позволяет добавить к аннотации новые термины. Он лишь исключает термины, для которых ряд параметров сходства анализируемой последовательности и гомологов не соответствует заданным критериям.

Метод логистической регрессии (LR – logistic regression) реализован во встроенном пакете stats (R Core Team, 2013), функция glm (family = binomial). Логистическая регрессия предсказывает вероятность принадлежности объекта к классу, например «спам» или «неспам». Он предсказывает вероятность принадлежности объекта к классу на основе взвешенной суммы признаков и пропускает ее через логистическую (сигмоидную) функцию, которая

нормализует результат в число (вероятность) от 0 до 1. Градиентный бустинг (XGB – eXtreme Gradient Boosting) использован в варианте, реализованном в пакете xgboost (Chen, Guestrin, 2016), функция xgb.train. Метод случайного леса (RF – random forest) применен в версии из пакета randomForest (Liaw, Wiener, 2002), функция randomForest. Как градиентный бустинг, так и метод случайного леса относятся к ансамблевым алгоритмам, основанным на деревьях решений (decision trees). Это означает, что итоговое предсказание – результат совокупной работы множества отдельных деревьев решений. Параметры алгоритмов градиентного бустинга и случайного леса указаны в табл. S12.

Параметры для моделей подбирали в процессе обучения, и в каждом методе их набор был одинаков для всех терминов GO, анализируемых последовательностей и их гомологов. Это термины, отражающие уровень сходства, аминокислотный состав и частоту встречаемости терминов GO (табл. S2). Если термин GO у гомолога присутствовал в аннотации анализируемой последовательности в обучающей выборке, значение функции предсказания в методе машинного обучения было равно 1, в противном случае – 0.

Для оценки точности методов машинного обучения использовали аминокислотные последовательности белков *Arabidopsis thaliana* и *Homo sapiens* (см. табл. 1). Набор этих последовательностей каждого из этих двух видов был разделен на две части: 80 % для обучения и 20 % для тестирования. Дополнительно была сформирована комбинированная выборка белков для организмов, представленных в табл. 1: из комбинированной выборки случайным образом для обучения были отобраны 50 000 последовательностей, а 20 000, не совпадающих с ними, – для тестирования моделей машинного обучения (табл. S3).

**Метрики качества.** Оценку точности аннотации проводили на языке R с использованием пакета dplyr (Wickham et al., 2025). Для этого были сформированы два списка: а) референсный список с аминокислотными последовательностями, аннотированными терминами GO из БД по модельным организмам, подробнее см. в табл. S1, и б) список, полученный путем функциональной аннотации с использованием различных методов аннотирования (см. рис. 1). Для оценки точности аннотации, полученной каждым из описанных выше методов, проводилось их сравнение с референсной аннотацией. В дальнейшем под True Positive (TP) мы понимаем термины GO, присутствующие в обоих списках; к False Positive (FP) относятся термины, присутствующие в списке предсказанной аннотации, но отсутствующие в референсном (истинном) списке; к False Negative (FN) относятся термины, присутствующие в референсном списке, но отсутствующие в списке предсказанной аннотации.

Для оценки аннотации белков использованы следующие метрики: Precision (PR), Recall (RC), Accuracy (AC), а также метрика F-score, которая была результирующей.

Precision (PR) – доля истинно положительных предсказаний среди всех положительных предсказаний метода:

$$\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100. \quad (1)$$

Recall (RC) – доля истинно положительных предсказаний среди всех истинных терминов в референсной аннотации:

$$\mathrm{RC} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100. \quad (2)$$

Accuracy (AC) представляет собой среднее арифметическое между Precision и Recall:

$$\mathrm{AC} = \frac{\mathrm{PR}+\mathrm{RC}}{2} \times 100. \quad (3)$$

F-score представляет собой гармоническое среднее между Precision и Recall. Эта метрика стремится к нулю, если значение Precision или Recall стремится к нулю:

$$F1 = 2\,\frac{\mathrm{PR}\times\mathrm{RC}}{\mathrm{PR}+\mathrm{RC}} \times 100. \quad (4)$$

Поскольку алгоритмы машинного обучения (LR, XGB, RF) оценивают вероятность принадлежности термина GO для анализируемой последовательности, а не бинарное решение, необходимо выбрать порог отсечения ($t$), по которому термин будет считаться предсказанным. Для учета несбалансированности данных и выбора оптимального порога, не зависящего от его конкретного значения, рассчитывалась метрика $F_{\max}$ для порога отсечения $t \in (0; 1)$ с шагом в 0.1. Термин GO считался верно предсказанным (положительным классом), если его прогнозируемая вероятность превышала порог $t$. $F_{\max}$ определяется как максимальное значение F-score($t$) по всем порогам:

$$F_{\max} = \max\left\{2\,\frac{\mathrm{PR}(t)\times\mathrm{RC}(t)}{\mathrm{PR}(t)+\mathrm{RC}(t)}\right\} \times 100. \quad (5)$$

В задачах прогнозирования GO-терминов, где распределение терминов по частоте встречаемости крайне не сбалансировано (некоторые термины очень часты, другие крайне редки), а классификация является многометочной (одному белку может соответствовать множество терминов), часто используется метрика $F_{\max}$. Она рассчитывается для всего набора предсказаний путем варьирования порога отсечения ($t$), начиная с которого предсказанный ML-моделью термин считается положительным. $F_{\max}$ показывает максимальное качество, которого может достичь модель в идеальном случае выбора порога. В отличие от $F1$, которая рассчитывается для фиксированного порога, $F_{\max}$ оценивает качество ранжирования терминов по вероятности.

**Сравнение с другими методами.** Для верификации разработанного нами метода OrthoML2GO сравнили его с методами Blast2GO (https://www.blast2go.com/) (Conesa et al., 2005) и PANNZER2 (http://ekhidna2.biocenter.helsinki.fi/sanspanz/) (Törönen et al., 2018). Поиск гомологов BLAST запускали на вычислительном комплексе центра коллективного пользования «Биоинформатика» ИЦиГ СО РАН. Параметры запуска Blast2GO и PANNZER2 были выбраны по умолчанию.

## Результаты и обсуждение

### Влияние информации об ортогруппах на точность предсказания терминов GO

Для оценки влияния информации об ортогруппах на точность предсказания функции было проведено сравнение

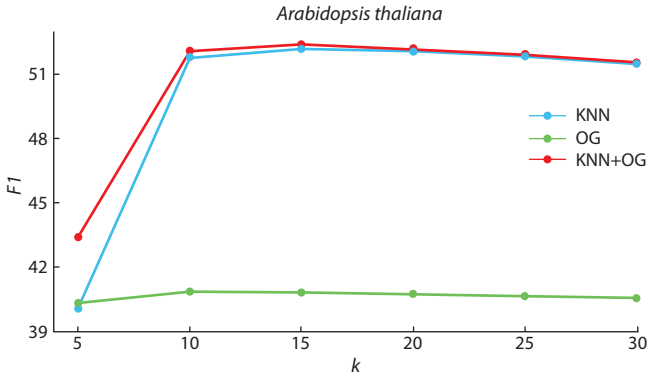


**Рис. 2.** Зависимость меры *F1* на белках *Arabidopsis thaliana* от параметра *k* (число ближайших гомологов) для трех вариантов аннотации.

По оси *X* отложены значения *k*, по оси *Y* – значения меры *F1*(%). Линии разного цвета соответствуют разным алгоритмам аннотации нашего метода: KNN – синий, OG – зеленый, KNN+OG – красный.

*F1* меры на трех методах аннотации терминами GO тремя алгоритмами (KNN, OG и KNN+OG) в зависимости от числа ближайших гомологов для последовательностей *Arabidopsis thaliana* (рис. 2).

Из рисунка 2 видно, что для всех трех вариантов аннотации наблюдается зависимость меры *F1* от числа *k*. Однако характер этих зависимостей различен: OG демонстрирует самую низкую точность (*F1* < 41 %). Для метода OG, как и для остальных методов, наблюдается максимум при *k* = 15. При этом дальнейшее увеличение параметра *k* приводит к монотонному, хотя и незначительному снижению меры *F1*. Мы полагаем, что для наиболее точного предсказания достаточно определить корректную ортологическую группу белка, которая идентифицируется уже при малых значениях *k*. Дальнейшее увеличение *k* приводит лишь к зашумлению предсказания из-за увеличения ложноположительных терминов GO от ортогрупп, к которым рассматриваемый белок на самом деле не относится.

Метод KNN демонстрирует выраженную зависимость точности от параметра *k*. При малых значениях (*k* = 5) мера *F1* самая низкая (~40 %) и ниже метода OG и KNN+OG, что, вероятно, связано с недостаточным количеством гомологов для надежного статистического вывода и высокой чувствительностью к шуму и возможным ошибкам аннотации отдельных последовательностей. При увеличении *k* до 15 наблюдается рост *F1* до максимального значения (~52 %), однако дальнейшее увеличение *k* приводит к постепенному снижению точности, поскольку в выборку начинают попадать отдаленные гомологи, несущие нерелевантную для целевой последовательности функциональную информацию (ложноположительные термины GO).

Объединение методов KNN и OG (KNN+OG) приводит к увеличению *F1*-меры при всех значениях параметра *k*, причем наибольший прирост (более 3 % по абсолютному значению) наблюдается именно при *k* = 5. Вероятно, этот эффект можно объяснить тем, что при малом *k* список гомологов может быть неустойчивым и статистически ненадежным. Добавление информации об ортогруппе, которая агрегирует данные о функции целой группы эволюционно родственных генов, стабилизирует предсказание и позволяет компенсировать недостаточность данных от малого числа ближайших соседей.

Стоит отметить, что значение *F1*-меры в диапазоне 40–52 % является конкурентоспособным результатом для задачи предсказания функций белков, что подтверждается сравнением с другими популярными методами (см. раздел «Сравнение точности методов KNN, KNN+OG и OrthoML2GO с инструментами Blast2GO и PANNZER2»). Это связано со сложной природой задачи: во-первых, как было показано ранее, аннотация GO является множественной, т. е. одному белку соответствует множество терминов, и предсказание считается верным, только если найдены все правильные термины и не добавлены лишние. Во-вторых, распределение терминов GO крайне не сбалансировано: некоторые термины очень часты, другие крайне редки, что дополнительно усложняет достижение высокой точности. Таким образом, абсолютное значение *F1*-меры следует интерпретировать в контексте сложности задачи и в сравнении с альтернативными подходами.

Для других организмов результаты применения трех подходов показаны в приложении (см. табл. S4–S9). Объединение методов KNN и OG (KNN+OG) позволяет получить интегральное предсказание, которое демонстрирует наибольший выигрыш в точности при малых значениях параметра *k* для всех организмов, кроме *Chlamydomonas reinhardtii*. Например, для белков *Danio rerio* при *k* = 5 метод KNN+OG превосходит базовый KNN более чем на 13 % по абсолютному значению *F1*-меры (74.66 против 61.37 %). Это объясняется тем, что при малом *k* список гомологов может быть статистически ненадежным и чувствительным к шуму в аннотациях отдельных последовательностей. Добавление информации об ортогруппе, которая агрегирует данные о функции целой группы эволюционно родственных генов, стабилизирует предсказание и позволяет компенсировать недостаточность данных от малого числа ближайших соседей. Следовательно, гибридный подход KNN+OG не только показывает лучший результат в пике (при *k* = 15), но и существенно снижает зависимость точности предсказания от параметра *k*, делая метод более устойчивым.

Таким образом, объединение вариантов KNN и OG (KNN+OG) позволяет получить интегральное предсказание, давая лучшую оценку по сравнению с каждым из них, для всех значений параметра *k* большинства организмов и для машинного обучения будет использован именно он.

### Верификация терминов GO различными алгоритмами машинного обучения

Для верификации ложноположительных терминов GO, полученных на предыдущем этапе, использованы такие алгоритмы машинного обучения, как логистическая регрессия (LR), градиентный бустинг (XGB) и случайного леса (RF) (см. раздел «Верификация терминов методами машинного обучения»). Сравнение точности методов машинного обучения с помощью меры $F_{max}$ (см. раздел «Сравнение с другими методами») на тестовых данных *Arabidopsis thaliana*, *Homo sapiens* и комбинированной выборки из 20000 последовательностей разных организмов представлено в табл. 2.

Логистическая регрессия демонстрирует значительно более низкие значения $F_{max}$ по сравнению с методами градиентного бустинга и случайного леса, причем разница может достигать более 25 %. Вероятно, это связано с тем, что ансамблевые методы (XGB и RF), в отличие от линейной модели LR, способны улавливать сложные нелинейные взаимоотношения между признаками. Кроме того, эти методы более устойчивы к шуму в данных за счет процедур бэггинга (RF) и бустинга (XGB), которые усредняют прогнозы множества отдельных деревьев решений, уменьшая влияние выбросов и некорректных аннотаций отдельных белков. Градиентный бустинг показывает лучшие результаты на последовательностях арабидопсиса и общей выборке всех организмов, но лишь незначительно уступает методу случайного леса на белках человека – разница в величине $F_{max}$ составляет всего 0.1 %. Таким образом, для финальной версии метода OrthoML2GO выбран метод машинного обучения градиентный бустинг (XGB), как показавший лучшие результаты на тестовых выборках.

### Сравнение точности методов KNN, KNN+OG и OrthoML2GO с инструментами Blast2GO и PANNZER2

Для всесторонней оценки эффективности разработанного метода было проведено сравнение его точности с двумя широко используемыми инструментами автоматической функциональной аннотации – Blast2GO и PANNZER2. Сравнение выполняли на трех тестовых наборах данных: отдельные протеомы *Arabidopsis thaliana* и *Homo sapiens*, а также комбинированная выборка, включающая последовательности всех организмов, перечисленных в табл. 1. В качестве результирующей метрики для методов, не использующих машинное обучение (KNN, KNN+OG, Blast2GO), применялась *F1*-мера, в то время как для OrthoML2GO и PANNZER2, выдающих вероятностную оценку, использовалась метрика $F_{max}$, позволяющая оценить максимально достижимое качество модели при идеальном выборе порога отсечения (табл. 3).

Анализ результатов демонстрирует, что разработанный метод OrthoML2GO, интегрирующий поиск гомологов, анализ ортогрупп и верификацию терминов GO с помощью градиентного бустинга, показал статистически значимое преимущество по точности над всеми сравниваемыми методами на всех тестовых выборках. Можно отметить, что разработанный метод OrthoML2GO на основе алгоритмов *k* ближайших соседей, ортогрупп и градиентного бустинга продемонстрировал статистически значимое преимущество по точности над всеми другими рассматриваемыми методами на всех тестовых выборках. Так, для

**Таблица 2.** Сравнение меры $F_{max}$ на тестовых данных для разных алгоритмов машинного обучения, %

| Организм | LR | XGB | RF |
|---|---|---|---|
| *Arabidopsis thaliana* | 53.20 | **68.95** | 66.86 |
| *Homo sapiens* | 71.92 | **83.92** | 84.02 |
| Комбинированная выборка | 52.25 | **79.55** | 78.32 |

*Arabidopsis thaliana* OrthoML2GO достиг $F_{max}$ = 68.95 %, что на 18.21 % превысило результат PANNZER2 и на 14.65 % – Blast2GO (*F1* = 54.30 %). На белках человека, по сравнению с PANNZER2, OrthoML2GO оказался существенно лучше – 83.92 против 75.14 %, а для метода Blast2GO значения *F1* = 54.95 %. На общей выборке всех организмов наблюдалось улучшение показателя F меры более чем на 30 % по сравнению со всеми другими методами.

Следует отметить, что гибридный подход KNN+OG, лежащий в основе OrthoML2GO, уже демонстрирует небольшое, но стабильное улучшение по сравнению с базовым KNN на всех выборках, что подтверждает полезность интеграции информации об ортогруппах. Однако основной выигрыш в точности обеспечивает этап верификации с помощью градиентного бустинга (XGB), который эффективно фильтрует ложноположительные предсказания, возникающие из-за шума в аннотации.

Одним из возможных факторов успеха метода OrthoML2GO является интеграция эволюционной информации из гомологов и ортогрупп БД OrthoDB с последующей верификацией терминов GO методом градиентного бустинга. В отличие от методов PANNZER2 и Blast2GO, наш метод учитывает информацию об ортогруппах и верифицирует термины GO с помощью ансамблей деревьев решений, адаптивно отбирая наиболее информативные признаки. В итоге это позволило снизить долю ложноположительных аннотаций и повысить точность от 8 % (на белковых последовательностях человека) до 30 % (на комбинированной выборке) по сравнению с аналогами.

Важно отметить, что в основе наших моделей машинного обучения лежит выборка последовательностей из БД OrthoDB. В отличие от этого, методы Blast2GO и PANNZER2 используют для своей работы более широкие наборы данных из базы UniProt, включающие как последовательности, так и аннотации. Поэтому нельзя исключить смещения в оценках точности, присущих двум этим методам в силу указанных обстоятельств.

**Таблица 3.** Сравнение методов KNN, KNN+OG, OrthoML2GO (XGB), PANNZER2 и Blast2GO на трех наборах данных, %

| Набор данных | KNN* | KNN+OG* | OrthoML2GO (XGB) | PANNZER2 | Blast2GO* |
|---|---|---|---|---|---|
| *Arabidopsis thaliana* | 51.54 | 51.68 | 68.95 | 50.74 | 54.30 |
| *Homo sapiens* | 71.72 | 72.18 | 83.92 | 75.14 | 54.95 |
| Комбинированная выборка | 47.29 | 47.35 | 79.55 | 49.14 | 42.11 |

* Приведено значение *F1*-меры; для OrthoML2GO и PANNZER2 – значение $F_{max}$.

**Таблица 4.** Сравнение точности предсказания функции словарей GO на комбинированной выборке, %

| Алгоритм | BP | MF | CC |
|---|---|---|---|
| LR | 50.9 | 48.5 | 56.8 |
| RF | 78.4 | 77.0 | 82.9 |
| XGB (OrthoML2GO) | 78.8 | 79.8 | 83.6 |

### Оценка точности идентификации терминов GO для разных словарей

Для проведения более детального анализа работы метода выполнен сравнительный анализ точности предсказания терминов GO для трех основных аспектов (словарей) Gene Ontology: Biological Process (BP, биологические процессы), Molecular Function (MF, молекулярные функции) и Cellular Component (CC, клеточные компоненты). Результаты оценки на комбинированной тестовой выборке для различных алгоритмов машинного обучения, использованных на этапе верификации, представлены в табл. 4.

Результаты показывают, что все алгоритмы машинного обучения демонстрируют схожую тенденцию: наивысшая точность предсказания достигается для аспекта Cellular Component (CC), затем следует Molecular Function (MF), и несколько ниже точность для Biological Process (BP). Это согласуется с общепринятым представлением в биоинформатике: предсказание клеточной локализации (CC) часто является наиболее простой задачей, так как оно сильно коррелирует с наличием специфических сигнальных пептидов и доменов. Предсказание молекулярной функции (MF) также в значительной степени зависит от консервативных функциональных доменов. В свою очередь, предсказание участия в биологических процессах (BP) наиболее проблематично, так как один и тот же белок может участвовать в нескольких процессах, а сами процессы определяются сложными взаимодействиями множества белков, что труднее установить исключительно из данных о гомологии и ортологии.

Метод XGB, выбранный для OrthoML2GO, показал лучшие результаты среди всех протестированных алгоритмов по всем трем аспектам, что дополнительно подтверждает корректность его выбора в качестве финального классификатора. Наши результаты по классификации GO сравнимы с оценками точности других методов, приведенными в литературных источниках (табл. 5). Сравнение выполнено с использованием метрики $F_{max}$ по отдельным аспектам GO: BP – биологические процессы, MF – молекулярные функции, CC – клеточные компоненты.

Метод OrthoML2GO (см. табл. 4) продемонстрировал конкурентоспособные результаты: 78.8 % (BP), 79.8 % (MF) и 83.6 % (CC) на выборке из 20000 последовательностей семи разнородных организмов – как растений, так и животных. При сравнении видно, что OrthoML2GO превосходит большинство исследуемых методов по всем аспектам. Однако PANNZER2 показал более высокие значения для MF (85.8 %) и CC (85.3 %), хотя на меньшей по объему и менее разнообразной выборке (5000 последовательностей из Swiss-Prot).

Следует отметить, что прямое количественное сопоставление с другими методами может быть осложнено методологическими различиями. Во-первых, тестовые выборки существенно разнятся: большинство методов использует БД UniProt/Swiss-Prot, тогда как наша комбинированная выборка включает как растения, так и животных, что может влиять на сравнимость результатов. Во-вторых, критически важна версия Gene Ontology: OrthoML2GO опирается на последнюю версию аннотации OrthoDB v 12.0 (GO 2025), что может вызывать сложности в сопоставлении метрик качества.

Для демонстрации применимости метода OrthoML2GO к слабо изученным организмам был аннотирован протеом зеленой водоросли *Ostreococcus lucimarinus* (табл. S10 и S11). Метод предсказал функции для 5273 из 7603 белковых последовательностей. Анализ выявил преобладание таких биологических процессов, как фосфорилирование (GO:0016310) и трансляция (GO:0006412). Среди молекулярных функций наиболее частыми оказались связывание АТФ (GO:0005524) и нуклеотидов (GO:0000166), а среди клеточных компонентов – мембрана (GO:0016020) и ядро (GO:0005634). Эти результаты демонстрируют способность метода аннотировать слабо изученные протеомы и выявлять функциональные профили, характерные для немодельных организмов.

## Заключение

Разработанный нами и представленный в этой работе метод предсказания функций белков OrthoML2GO, интегрирующий поиск гомологов и ортологических групп в базе данных OrthoDB с градиентным бустингом, продемонстрировал высокую эффективность на тестовых выборках.

**Таблица 5.** Оценки точности аннотации терминами GO для различных словарей разными методами по литературным данным, %

| Метод | BP | MF | CC | Литературный источник |
|---|---|---|---|---|
| PANNZER2 | 78.4 | 85.8 | 85.3 | Törönen et al., 2018 |
| DeepGOPlus | 58.5 | 47.4 | 69.9 | Kulmanov, Hoehndorf, 2020 |
| GOLabeler | 58.6 | 37.2 | 69.1 | You et al., 2018 |
| NetGO 2.0 | 66.6 | 36.6 | 66.3 | Yao et al., 2021 |
| TALE+ | 66.7 | 45.9 | 67.7 | Cao, Shen, 2021 |

Одним из основных результатов является значительное повышение точности аннотации за счет комбинированного подхода, объединяющего метод *k* ближайших гомологов и информацию об ортологических группах (KNN+OG). Этот гибридный метод превзошел отдельные подходы KNN и OG, особенно при малых значениях параметра *k*. Верификация терминов GO с помощью алгоритмов машинного обучения, в особенности градиентного бустинга (XGB), позволила дополнительно повысить точность за счет эффективного отсева ложноположительных предсказаний, возникающих при использовании отдаленных гомологов и ортогрупп.

Полученные результаты подтверждают, что использование эволюционной информации, заключенной в ортогруппах OrthoDB, в сочетании с алгоритмами машинного обучения – эффективная стратегия для автоматического предсказания функций белковых последовательностей. Предложенный метод OrthoML2GO может стать хорошей альтернативой уже существующим методам. Дальнейшее улучшение точности возможно за счет оптимизации параметров машинного обучения, а также включения дополнительных источников биологической информации. В качестве перспективных исследований намечены следующие направления: оценка переносимости модели на слабо аннотированные протеомы и сравнительный анализ с другими методами, использующими машинное обучение, в том числе нейросетевыми.

## Список литературы / References

Altenhoff A.M., Glover N.M., Dessimoz C. Inferring orthology and paralogy. *Methods Mol Biol*. 2019;1910:149-175. doi 10.1007/978-1-4939-9074-0_5

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi 10.1016/S0022-2836(05)80360-2

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., … Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25-29. doi 10.1038/75556

Benso A., Di Carlo S., Ur Rehman H., Politano G., Savino A., Suravajhala P. A combined approach for genome wide protein function annotation/prediction. *Proteome Sci*. 2013;11(Suppl. 1):S1. doi 10.1186/1477-5956-11-S1-S1

Bradford Y.M., Van Slyke C.E., Ruzicka L., Singer A., Eagle A., Fashena D., Howe D.G., Frazer K., Martin R., Paddock H., Pich C., Ramachandran S., Westerfield M. Zebrafish information network, the knowledgebase for *Danio rerio* research. *Genetics*. 2022;220(4):iyac016. doi 10.1093/genetics/iyac016

Buchfink B., Xie C., Huson D.H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12(1):59-60. doi 10.1038/nmeth.3176

Cao Y., Shen Y. TALE: Transformer-based protein function Annotation with joint sequence-Label Embedding. *Bioinformatics*. 2021;37(18):2825-2833. doi 10.1093/bioinformatics/btab198

Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. In: KDD ‘16. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2016;785-794. doi 10.1145/2939672.2939785

Cheng S., Melkonian M., Smith S.A., Brockington S., Archibald J.M., Delaux P.M., Li F.W., … Graham S.W., Soltis P.S., Liu X., Xu X., Wong G.K. 10KP: A phylodiverse genome sequencing plan. *Gigascience*. 2018;7(3):1-9. doi 10.1093/gigascience/giy013

Conesa A., Götz S., García-Gómez J.M., Terol J., Talón M., Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005;21(18):3674-3676. doi 10.1093/bioinformatics/bti610

Dongardive J., Abraham S. Protein Sequence Classification Based on N-Gram and K-Nearest Neighbor Algorithm. In: Behera H., Mohapatra D. (Eds). Computational Intelligence in Data Mining. Vol. 2. Advances in Intelligent Systems and Computing. Vol. 411. Springer, New Delhi, 2016;163-171 doi 10.1007/978-81-322-2731-1_15

du Plessis L., Skunca N., Dessimoz C. The what, where, how and why of gene ontology – a primer for bioinformaticians. *Brief Bioinform*. 2011;12(6):723-735. doi 10.1093/bib/bbr002

Edgar R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460-2461. doi 10.1093/bioinformatics/btq461

Eisenberg D., Marcotte E.M., Xenarios I., Yeates T.O. Protein function in the post-genomic era. *Nature*. 2000;405(6788):823-826. doi 10.1038/35015694

Fitch W.M. Distinguishing homologous from analogous proteins. *Syst Biol*. 1970;19(2):99-113. doi 10.2307/2412448

Fitch W.M. Homology a personal view on some of the problems. *Trends Genet*. 2000;16(5):227-231. doi 10.1016/s0168-9525(00)02005-9

Galperin M.Y., Koonin E.V. From complete genome sequence to ‘complete’ understanding? *Trends Biotechnol*. 2010;28(8):398-406. doi 10.1016/j.tibtech.2010.05.006

Gene Ontology Consortium; Aleksander S.A., Balhoff J., Carbon S., Cherry J.M., Drabkin H.J., Ebert D., ... Ponferrada V., Zorn A., Ramachandran S., Ruzicka L., Westerfield M. The Gene Ontology knowledgebase in 2023. *Genetics*. 2023;224(1):iyad031. doi 10.1093/genetics/iyad031

Goodwin S., McPherson J.D., McCombie W.R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333-351. doi 10.1038/nrg.2016.49

Grigoriev I.V., Hayes R.D., Calhoun S., Kamel B., Wang A., Ahrendt S., Dusheyko S., Nikitin R., Mondo S.J., Salamov A., Shabalov I., Kuo A. PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res*. 2021;49(D1):1004-1011. doi 10.1093/nar/gkaa898

Hamilton J.P., Brose J., Buell C.R. SpudDB: a database for accessing potato genomic data. *Genetics*. 2025a;229(3):iyae205. doi 10.1093/genetics/iyae205

Hamilton J.P., Li C., Buell C.R. The rice genome annotation project: an updated database for mining the rice genome. *Nucleic Acids Res*. 2025b;53(1):1614-1622. doi 10.1093/nar/gkae1061

Huntley R.P., Sawford T., Mutowo-Meullenet P., Shypitsyna A., Bonilla C., Martin M.J., O’Donovan C. The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Res*. 2015;43(D1):1057-1063. doi 10.1093/nar/gku1113

Jensen L.J., Julien P., Kuhn M., von Mering C., Muller J., Doerks T., Bork P. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*. 2008;36(Database issue):250-254. doi 10.1093/nar/gkm796

Kharsikar S., Mugler D., Sheffer D., Moore F., Duan Z.H. A weighted k-nearest neighbor method for gene ontology based protein function prediction. In: Proceedings of the Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS ‘07). IEEE Computer Society, USA, 2007;25-31. doi 10.1109/IMSCCS.2007.13

Kriventseva E.V., Rahman N., Espinosa O., Zdobnov E.M. OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res*. 2008;36(Database issue):271-275. doi 10.1093/nar/gkm845

Kulmanov M., Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*. 2020;36(2):422-429. doi 10.1093/bioinformatics/btz595

Kuzniar A., van Ham R.C., Pongor S., Leunissen J.A. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*. 2008;24(11):539-551. doi 10.1016/j.tig.2008.08.009

Lewin H.A., Robinson G.E., Kress W.J., Baker W.J., Coddington J., Crandall K.A., Durbin R., …van Sluys M.A., Soltis P.S., Xu X.,

Yang H., Zhang G. Earth BioGenome project: Sequencing life for the future of life. *Proc Natl Acad Sci USA*. 2018;115(17):4325-4333. doi 10.1073/pnas.1720115115

Liaw A., Wiener M. Classification and Regression by randomForest. *R News*. 2002;2(3):18-22. doi 10.32614/CRAN.package.randomForest

Öztürk-Çolak A., Marygold S.J., Antonazzo G., Attrill H., Goutte-Gattat D., Jenkins V.K., Matthews B.B., Millburn G., Dos Santos G., Tabone C.J.; FlyBase Consortium. FlyBase: updates to the *Drosophila* genes and genomes database. *Genetics*. 2024;227(1):iyad211. doi 10.1093/genetics/iyad211

Pearson W.R. An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics*. 2013;42(3):3.1.1-3.1.8. doi 10.1002/0471250953.bi0301s42

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, 2013. Available: http://www.R-project.org/

Reiser L., Bakker E., Subramaniam S., Chen X., Sawant S., Khosa K., Prithvi T., Berardini T.Z. The Arabidopsis Information Resource in 2024. *Genetics*. 2024;227(1):iyae027. doi 10.1093/genetics/iyae027

Sanderson T., Bileschi M.L., Belanger D., Colwell L.J. ProteInfer, deep neural networks for protein functional inference. *eLife*. 2023;12:e80942. doi 10.7554/eLife.80942

Steinegger M., Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. 2017;35(11):1026-1028. doi 10.1038/nbt.3988

Suzuki S., Kakuta M., Ishida T., Akiyama Y. GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PLoS One*. 2014;9(8):e103833. doi 10.1371/journal.pone.0103833

Tegenfeldt F., Kuznetsov D., Manni M., Berkeley M., Zdobnov E.M., Kriventseva E.V. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res*. 2025;53(D1):D516-D522. doi 10.1093/nar/gkae987

Törönen P., Medlar A., Holm L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res*. 2018;46(W1):W84-W88. doi 10.1093/nar/gky350

Wickham H., François R., Henry L., Müller K., Vaughan D. dplyr: A Grammar of Data Manipulation. R package version 1.1.4. 2025. doi 10.32614/CRAN.package.dplyr

Yao S., You R., Wang S., Xiong Y., Huang X., Zhu S. NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic Acids Res*. 2021;49(W1):W469-W475. doi 10.1093/nar/gkab398

You R., Zhang Z., Xiong Y., Sun F., Mamitsuka H., Zhu S. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*. 2018;34(14):2465-2473. doi 10.1093/bioinformatics/bty130

Yuan Q., Xie J., Xie J., Zhao H., Yang Y. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. *Brief Bioinform*. 2023;24(3):bbad117. doi 10.1093/bib/bbad117

**Конфликт интересов.** Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 24.07.2025. После доработки 10.09.2025. Принята к публикации 15.09.2025.