1. **Аннотация**

С помощью БД OrthoDB и с использованием программы USEARCH нами был разработан метод аннотации функций генов белковых последовательностей, который при k =4 ближайших гомологов имеет высокую точность как для молодых генов, так и для более старых генов. Верификация нашего метода с классической и широко используемой программой Blast2GO показала, что помимо высокой скорости аннотации предложенный нами метод по точности превосходит Blast2GO на величину до 18%.

2. **Тема работы:**

«Компьютерная аннотация белковых последовательностей растений на основе гомологии»

3. **Состав коллектива:**

Афонников Дмитрий Аркадьевич, к.б.н., в.н.с. Института Цитологии и Генетики СО РАН, ada@bionet.nsc.ru

Генаев Михаил Александрович, к.б.н., с.н.с. Института Цитологии и Генетики СО РАН, mag@bionet.nsc.ru

Пронозин Артем Юрьевич, м.н.с. Института Цитологии и Генетики СО РАН, PronozinAU@bionet.nsc.ru

4. **Информация о гранте:**

Разработка алгоритмов финансировалась Курчатовским геномным центром Института Цитологии и Генетики СО РАН, соглашение с Министерством образования и науки Российской Федерации № 075-15-2019-1662

5. **Научное содержание работы:**

5.1 Постановка задачи

Разработка метода аннотации функций генов растений на основе поиска гомологов в базах белковых последовательностях с использованием программы USEARCH, имеющего высокую точность для генов различных возрастов и его сравнение с существующими алгоритмами

5.2 Современное состояние проблемы

С появлением методов секвенирования нового поколения число последовательностей ДНК, РНК и белков растет огромными темпами. Источниками поступления такого большого количества данных являются различные геномные, метагеномные и транскриптомные проекты. В результате выполнения таких проектов прочитывается большое количество белковых последовательностей, функция которых оказывается неизвестна. Без знания функции белков невозможно интерпретировать результат изменения экспрессии генов для жизнедеятельности клеток, определять эффект мутаций в геномах организмов. Поэтому необходимы биоинформатические методы, основанные лишь на информации о последовательности белка, которые позволяют

определить его молекулярную функцию, роль в клеточных процессах и клеточную локализацию. Эта информация может быть описана в терминах генной онтологии, которые представлены в базе данных Gene Ontology (GO).

Аннотацией или предсказанием функций белка называют определение биологической роли белка и его значения в процессе функционирования клетки. Одним из биоинформатических способов предсказания функции белка являются методы, основанные на гомологии. Так как белки, имеющие сходные последовательности, как правило, являются гомологичными и имеют сходную функцию. Поэтому термины, которыми описана функция известного белка можно присвоить последовательности белка с неизвестной функцией, если они гомологичны (обе последовательности имеют высокое сходство).

В процессе эволюции некоторые гены дуплицируются и после этого могут изменить свою функцию. Это приводит к образованию нового семейства генов. Семейства генов, которые образовались на ранних этапах эволюции живых организмов считаются древними, а образовавшиеся относительно недавно, на этапе формирования современных видов, считаются молодыми. Для молодых генов характерно отсутствие большого числа гомологов у других видов. Очевидно, что если гомологичные последовательности для исходной последовательности отсутствуют, то это не позволяет получить информацию для ее аннотации. Такая ситуация характерна для молодых или таксон-специфичных генов, представленных в одном или нескольких ближайших видах и отсутствующих у других организмов. Таким образом, возраст гена может оказывать сильное влияние на точность аннотации.

5.3 Подробное описание работы, включая используемые алгоритмы

Материалом послужили последовательности белок-кодирующих генов в формате FASTA. Белковые последовательности *Arabidopsis thaliana* были взяты из базы данных TAIR. Каждому гену из генома *A.thaliana* соответствовала одна аминокислотная последовательность. Всего мы использовали для анализа 27655 аминокислотных последовательностей.

Для оценки возрастов генов *A.thaliana* использовались данные из работы [Phylostratigraphic analysis shows the earliest origination of the abiotic stress associated genes in A. thaliana], в которой для каждого гена *A.thaliana* рассчитана величина PAI. В процессе оценки влияния возраста гена на эффективность его аннотации мы сгруппировали 15 возрастов генов в три большие группы: старые (0<=PAI<=7), средние (7<=PAI<=12), молодые (13<=PAI<=17).

Локальное выравнивание и поиск гомологов проводилось с помощью алгоритма USEARCH v 11 (Usearch_local). Для поиска ортологичных групп генов мы использовали базу данных OrthoDB v 10.0. База данных включает аннотацию GO для части последовательностей и, таким образом, является удобным источником их классификации на ортологи и аннотации GO. Кроме того, эта база данных предоставляет классификацию белковых последовательностей на ортологические семейства, для которых также представлена обобщенная аннотация функций белка в терминах GO.

В процессе своей работы мы разработали 3 метода аннотации:

I. **На основе *k* ближайших гомологов (KNN)**. Это базовый метод, который соответствует методу, описанному в работе [32] Hayai-Annotation Plants, за исключением того, что мы используем для поиска гомологов БД OrthoDB. Для каждой искомой последовательности находим *k* ближайших гомологов по уровню сходства, которые выдаются в результате поиска по БД OrthoDB программой Usearch с параметрами по умолчанию. Искомой последовательности присваиваются термины GO *k* наиболее сходных последовательностей, приведенные в БД OrthoDB (Рисунок 1);

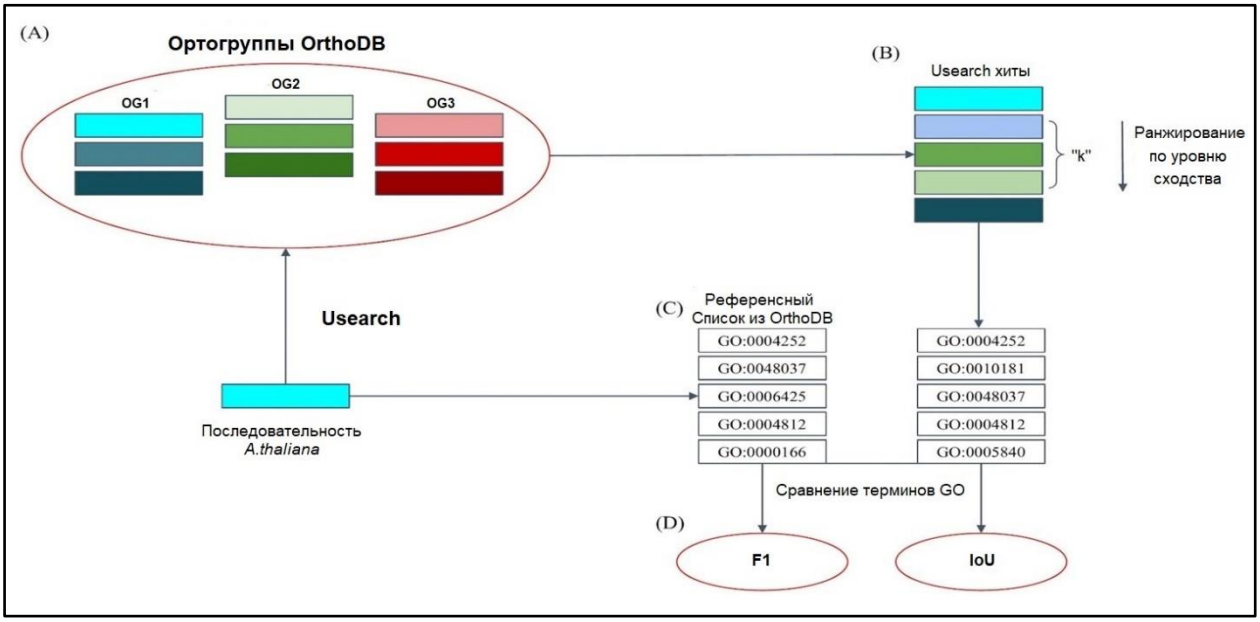

Рисунок 1 – Схема базового метода предсказания функции: k ближайших гомологов (KNN)

II. **С помощью аннотации ортологических групп (OG)**. Этот метод опирается на концепцию ортологии: ортологичные последовательности из разных видов выполняют одинаковые функции. Для семейств ортологов БД OrthoDB предоставляет аннотацию функции терминами GO, которую, следовательно, можно использовать для предсказания функции искомого гена. Для каждой искомой последовательности на основе *k* ее ближайших гомологов методом голосования определяется наиболее представленные ортогруппы (их может быть несколько из-за иерархической структуры ортогрупп OrthoDB), искомой последовательности присваиваются термины GO этих ортогрупп (Рисунок 2);

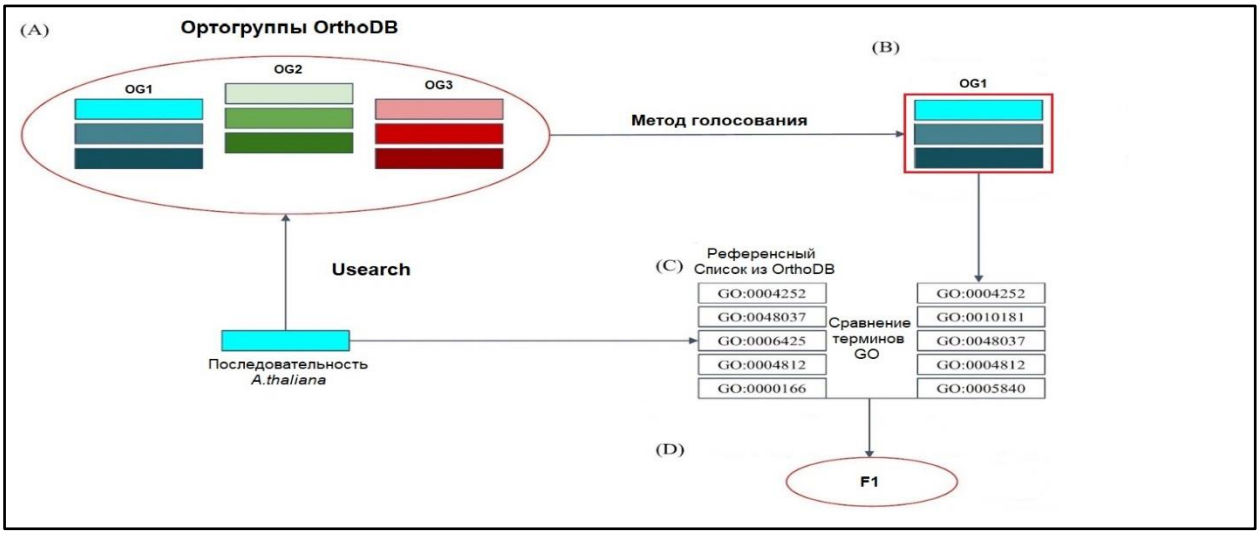

Рисунок 2 – Схема метода предсказания функции на основе ортологических групп (OG)

III. **Комбинация двух предыдущих методов (KNN+OG)**. Этот метод основан на объединении терминов GO для последовательности, полученных как в результате поиска k ближайших гомологов и метода с учетом поиска ортологов.

В процессе оценки функциональной аннотации были сформированы два списка: Референсный список с белковыми последовательностями и терминами GO из БД OrthoDB и список, который был получен путем функциональной аннотации с использованием различных методов аннотации.

Таким образом под True Positive (TP) мы понимаем те термины GO которые есть в обоих списках; к False Positive (FP) относятся термины, которые есть в списке с функциональной аннотацией и отсутствует в референсном; и под False Negative (FN) мы понимаем те термины, которые есть в референсном списке, но отсутствуют в списке с аннотацией

Для оценки аннотации белков были использованы следующие метрики: Специфичность (SP), Чувствительность (SN), Точность (AC) и F1-мера. Сравнение было выполнено с использованием всех аннотированных генов, связанных с его терминами GO, извлеченными из БД OrthoDB.

Специфичность (SP) – процент результатов, который соответствует исследуемой выборке.

$$SP = \frac{TP}{TP + FP} * 100$$

Чувствительность (SN) – процент совпавших результатов с исследуемой выборкой, правильно классифицированных.

$$SN = \frac{TP}{TP + FN} * 100$$

Точность (AC) представляет собой среднее арифметическое между специфичностью и чувствительностью

$$AC = \frac{SN + SP}{2} * 100$$

F1 мера представляет собой гармоническое среднее между специфичностью и чувствительностью. Она стремится к нулю, если специфичность или чувствительность стремится к нулю

$$F1 = 2\frac{SP * SN}{SP + SN} * 100$$

5.4 Полученные результаты

Нами было проведено сравнение меры F1 для трех методов предсказания функции (Рисунок 3).

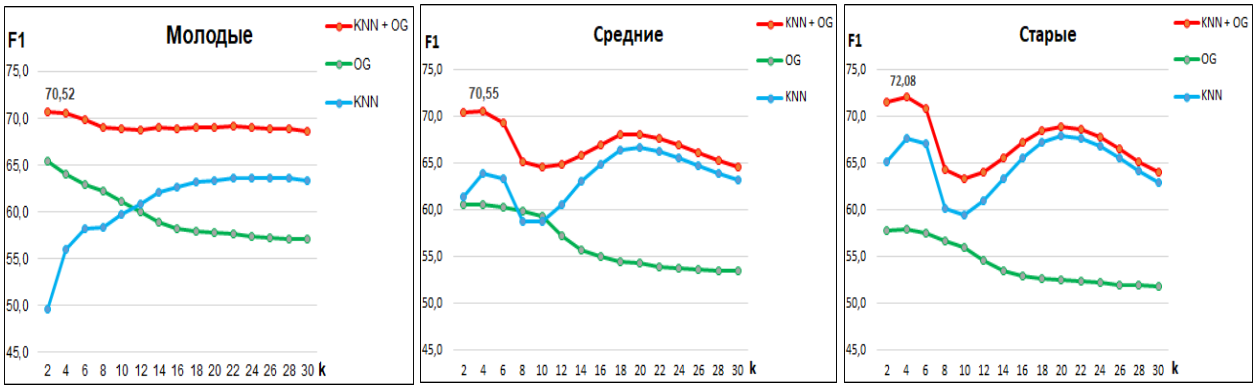

Рисунок 3 - Зависимость F1 меры генов разных возрастов от k ближайших гомологов для разных методов аннотации

В результате анализа можно сделать вывод, что объединение результатов KNN и OG дает принципиальное улучшение точности распознавания функций для генов всех возрастов и при k=4 позволяет нивелировать эффект возраста генов на точность их аннотирования, вне зависимости от возраста точность становится примерно одинаковой (F1=70,5-72). Для верификации этого метода мы решили сравнить его с алгоритмом Blast2GO (Таблица 1)

Таблица 1 – Сравнение точности метода KNN+OG с Blast2GO для генов разных возрастов

	KNN+OG				**Blast2GO**			
	SN	SP	AC	F1	SN	SP	AC	F1
Молодые	61.49	82.66	72.08	**70.52**	42.55	66.41	54.48	**51.86**
Средние	58.22	89.50	73.86	**70.55**	46.02	72.97	59.50	**56.44**
Старые	58.94	92.75	75.85	**72.08**	46.39	76.86	61.62	**57.85**

Таким образом можно сделать вывод что в среднем наш метод точнее Blast2GO на 14-18% для генов разных возрастов. Также стоит отметить, что среднее время работы нашего алгоритма ~ 2,5 часа на 8 ядрах процессора, а среднее время работы программы Blast2GO (запуск программы blastp) ~ 588 часов (25 дней) на 16 ядрах процессора.

6. Эффект от использования кластера в осуществлении цели:

Хранение БД OrthoDB и поиск по ней с помощью программы USEARCH был полностью реализован на информационно-вычислительном кластере Новосибирского Государственного Университета (Узел HP XL230a Gen9; 24 ядра и 192 ГБ ОЗУ). Таким образом использование ресурсов ИВЦ НГУ является определяющим для успешного достижения целей нашей работы.