МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ЕСТЕСТВЕННЫХ НАУК
КАФЕДРА ИНФОРМАЦИОННОЙ БИОЛОГИИ

КОМПЬЮТЕРНАЯ АННОТАЦИЯ БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ РАСТЕНИЙ НА ОСНОВЕ ГОМОЛОГИИ

Novosibirsk State University
*THE REAL SCIENCE

Автор работы: Студент группы № 20426 Малюгин Е. В.
Научный руководитель: Канд. биол. наук, Афонников Д. А.

Новосибирск – 2022

Задача аннотации функций белковых последовательностей

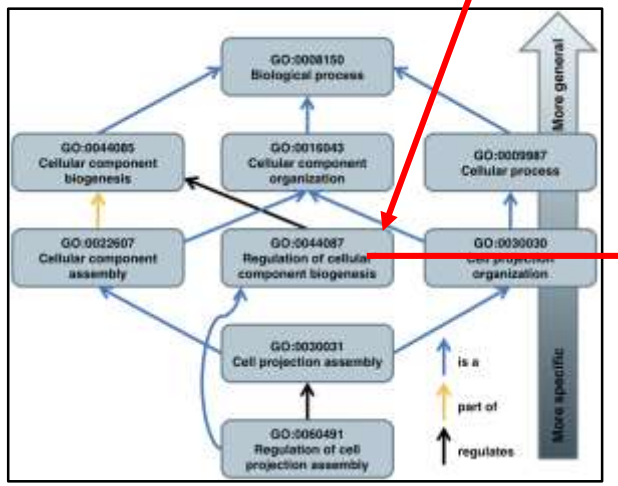
- Число секвенированных последовательностей растет огромными темпами и требует эффективного предсказания функций белков.

- Для описания функций используется база данных Gene Ontology, в которой описаны функции миллионов генов стандартными терминами.

- Для аннотации новых генов используется принцип гомологии: сходные последовательности выполняют сходные функции.

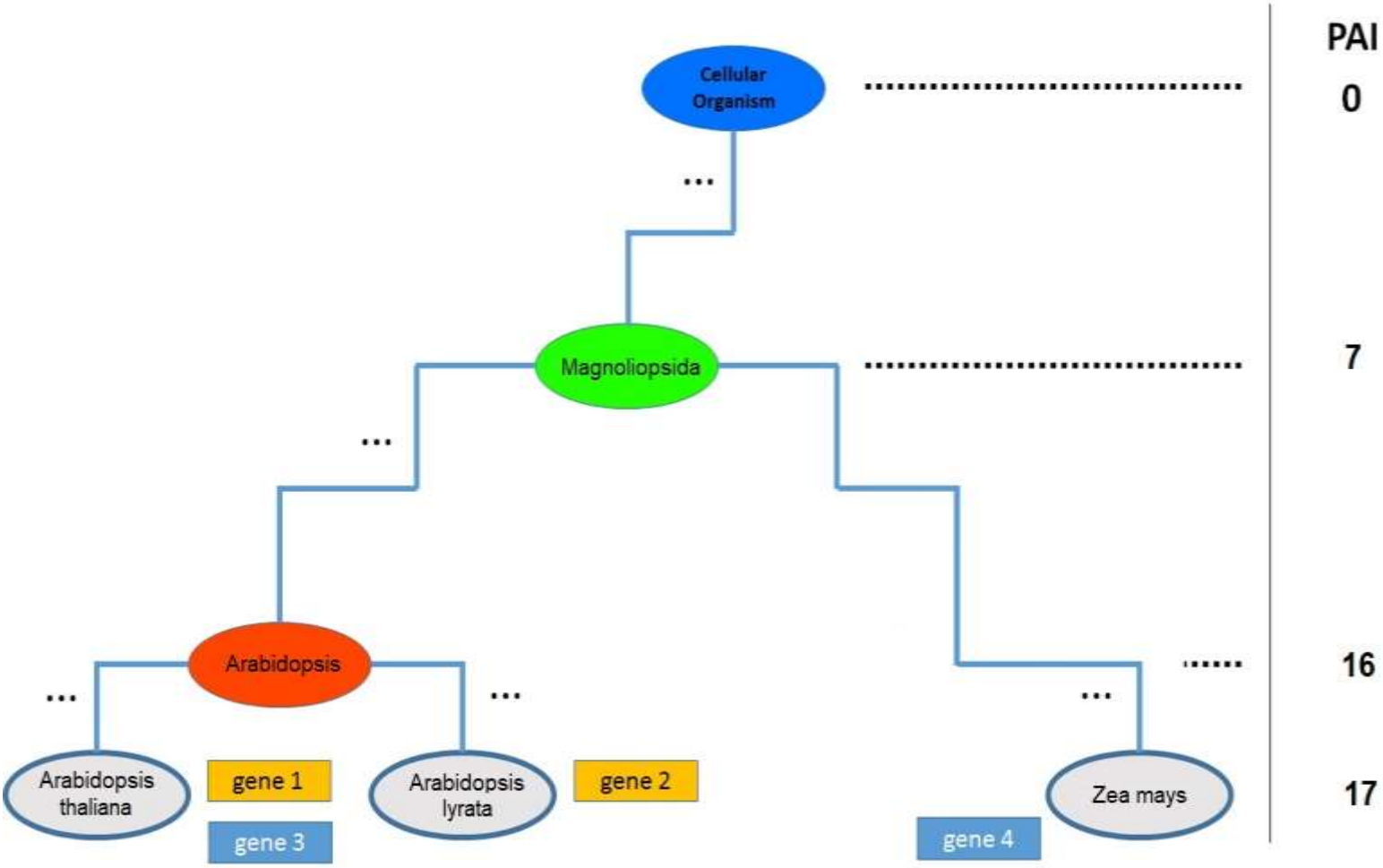
Поиск гомологов Blastp

БД Gene Ontology



Присваивание терминов GO для неизвестного гена на основе сходства

Ортологи и возраст генов



Одинаковый цвет=Одинаковая функция (ортология)

PAI(gene1) = PAI(gene2) = 16 («Arabidopsis» – молодой ген)

PAI(gene3) = PAI(gene4) = 7 («Magnoliopsida» – “ген среднего возраста”)

Возраст гена – определяется как таксономическое положение наиболее позднего общего предка, в котором встречаются его ортологи (гомологи с одинаковой функцией)

Древние гены: имеют много гомологов в базах данных, как правило хорошо аннотированы.

Молодые гены (орфанные) – мало гомологов (1-3 шт.), слабо аннотированы, как правило это гены с неизвестной функцией.

Необходимо исследование влияния возраста гена на точность аннотации его функции

Цели и задачи

Целью работы является создание метода аннотации функций генов растений на основе поиска гомологов в базах белковых последовательностях, имеющего высокую точность для генов различных возрастов.

Задачи:

1. Разработка конвейера для анализа и предсказания функций генов растений по гомологии на основе k ближайших гомологов из базы данных OrthoDB, и оценка его точности на основе аннотации генов *Arabidopsis thaliana*;
2. Оценка влияния возраста генов на точность аннотации их функции;
3. Разработка метода предсказания функции генов с учетом аннотации ортологических групп из БД OrthoDB и сравнение его точности с существующими методами;
4. Оценка точности метода для генов из 5 дополнительных видов растений, представляющих разные таксоны.

Последовательности и базы данных

Использовались:
27 655 **последовательностей белок-кодирующих генов *Arabidopsis thaliana*** из БД TAIR

Последовательности 5 дополнительных видов (*Chlamydomonas reinhardtii*, *Oryza sativa*, *Zea mays*, *Solanum lycopersicum*, *Solanum tuberosum*) из БД KEGG.

Возраста генов оценивались программой Orthoscape (Мустафиным З.С.), классификация возрастов проводилась на 3 категории: старые (*Cellular organisms-Tracheophyta*), средние (*Magnoliophyta-Malvids*) и молодые (*Brassicales-Arabidopsis thaliana*) (**Для *A.thaliana***)

В качестве источника аннотации белковых последовательностей использовалась БД OrthoDB v 10.0

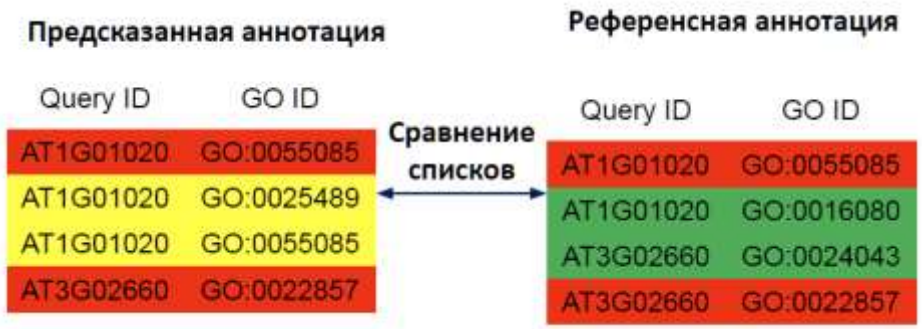
Поиск гомологов осуществлялся программой USEARCH v 11 (алгоритм usearch_local)

Для сравнения точности предсказания функции использовалась программа Blast2GO.

Программы для анализа данных были реализованы на языке R.

Оценка качества предсказания функций генов

Расчет параметров TP, TN, FP на основе совпадений в списках аннотаций:

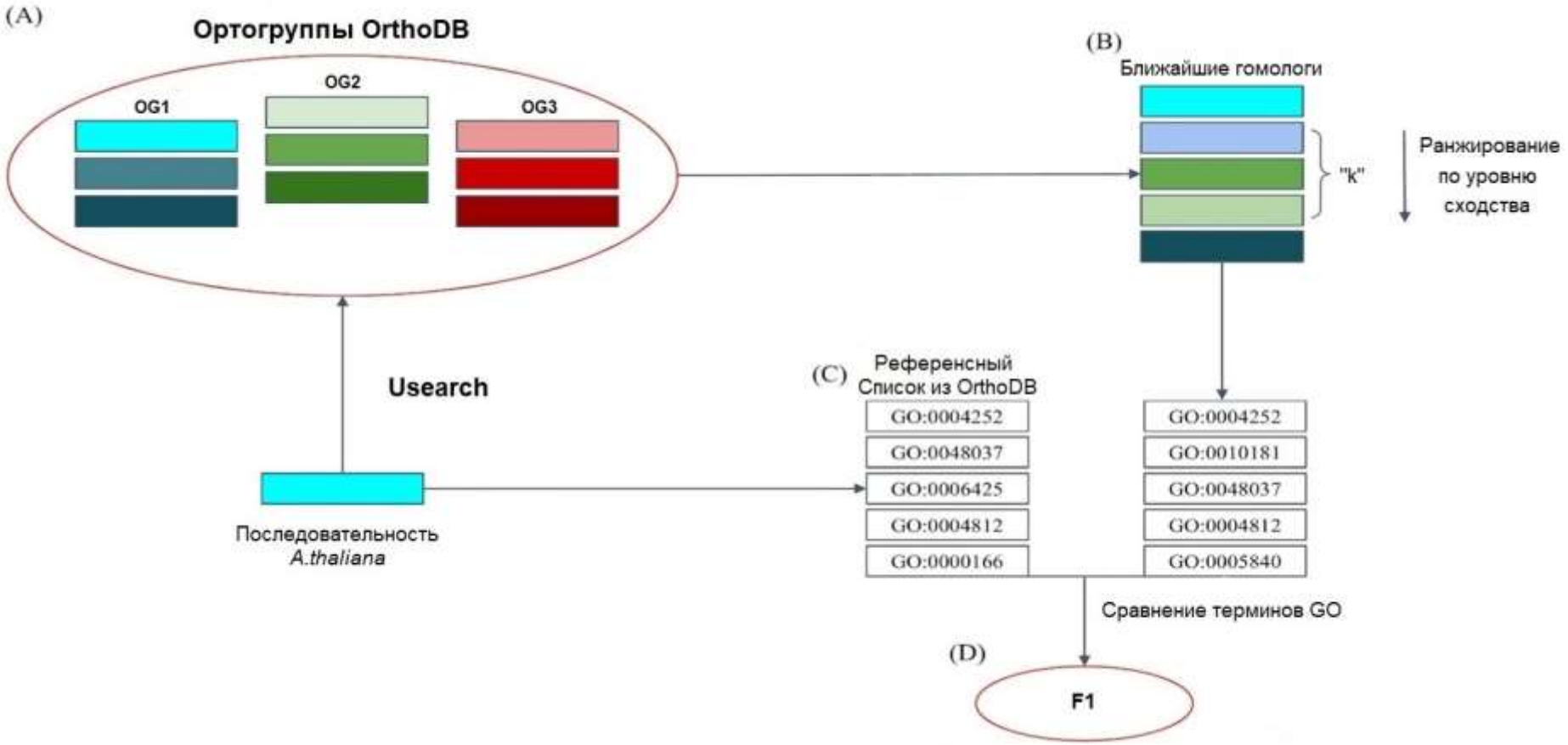


Красным цветом – TP (термин предсказан правильно)
Желтым цветом – FP (предсказанный термин отсутствует в аннотации)
Зеленым цветом – FN (аннотированный термин не предсказан)

Расчет параметров точности аннотации на основе TP, TN, FP:

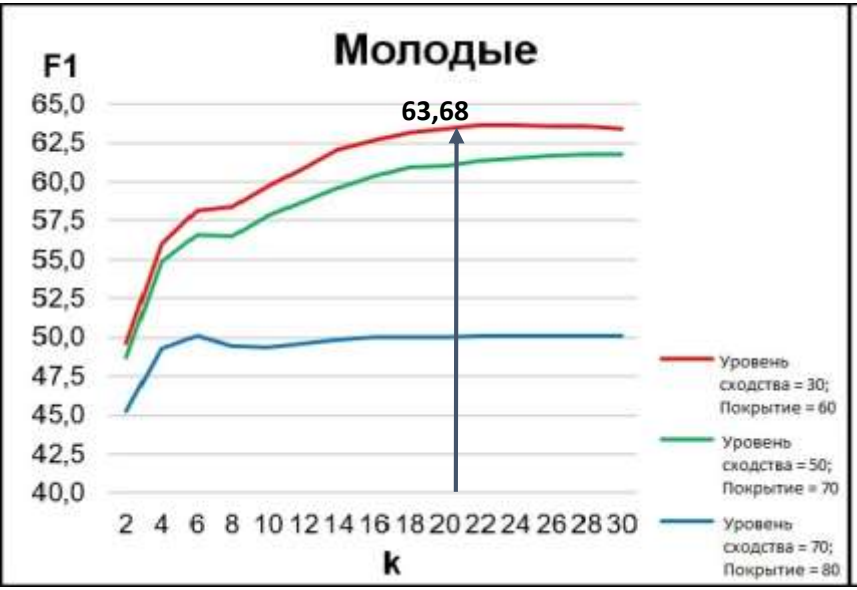
1. **Специфичность** (SP): $SP = \frac{TP}{TP+FP} * 100$; = 100 если FP=0
2. **Чувствительность** (SN): $SN = \frac{TP}{TP+FN} * 100$; = 100, если FN=0
3. **Точность** (AC): $AC = \frac{SN+SP}{2} * 100$; = 100, если нет ошибок предсказания
4. **F1-мера**: $F1 = 2\frac{SP*SN}{SP+SN} * 100$; = 100 , если нет ошибок предсказания

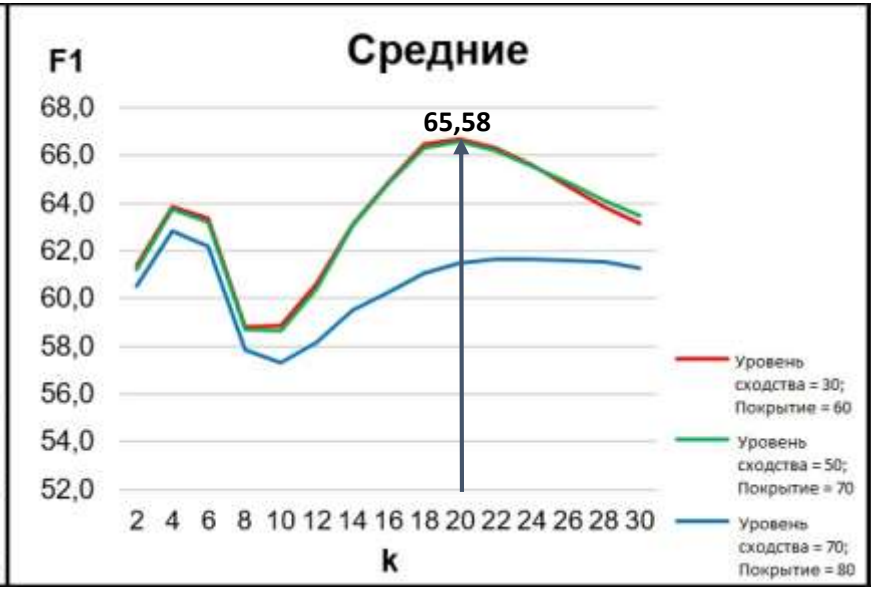
Базовый метод предсказания функции: k ближайших гомологов (KNN)

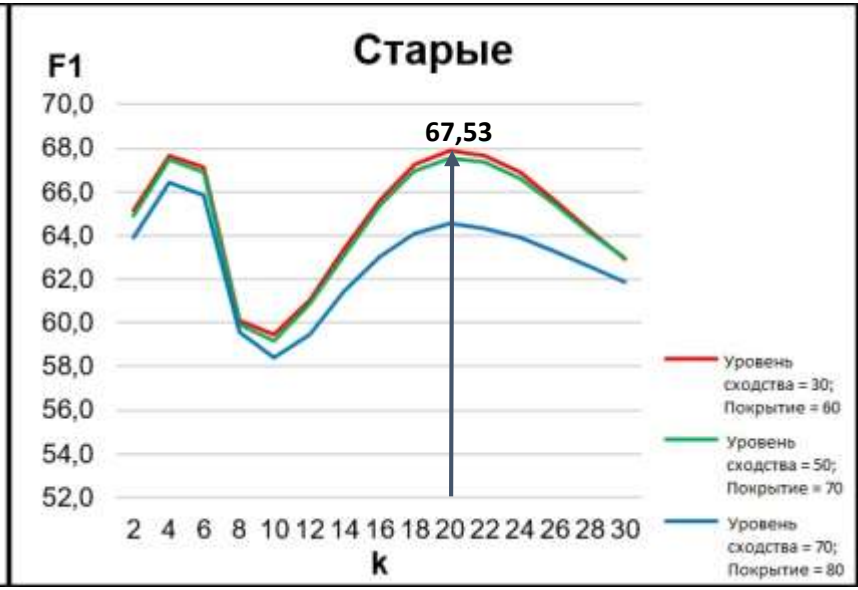


Для каждой искомой **последовательности находим *k* ближайших гомологов** по уровню сходства, которые выдаются в результате поиска по БД OrthoDB программой Usearch.
Искомой последовательности **присваиваются термины GO *k* наиболее сходных последовательностей**

Точность предсказания функций в зависимости от сходства, покрытия и числа гомологов (k) для генов разных возрастов



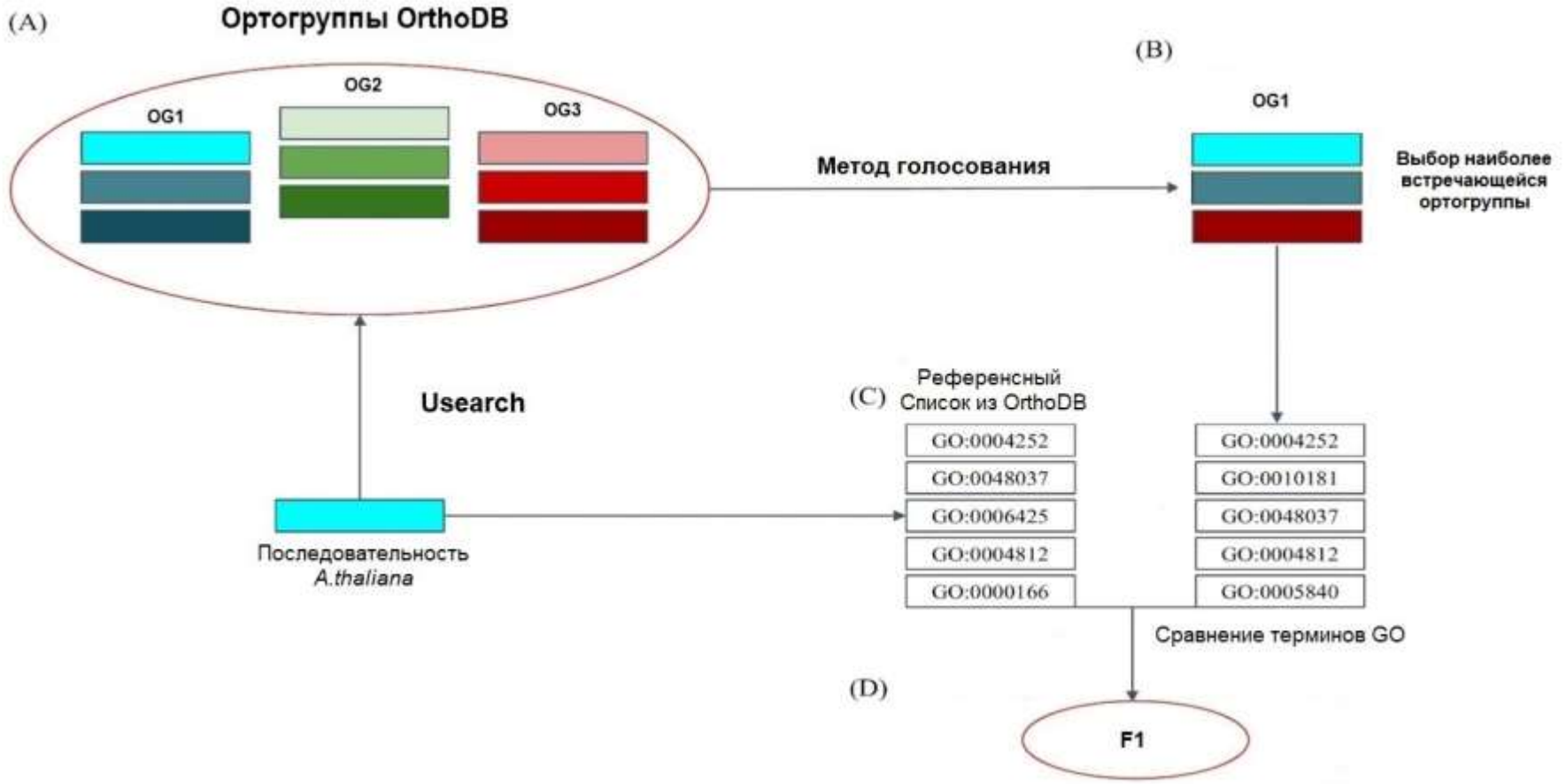




Наилучшая точность достигается для **порога сходства** последовательностей **30%**, **покрытия** выравниванием – **60%** и **k = 20** ближайших гомологов

Оптимальные значения параметров для генов разных возрастов оказались одинаковыми

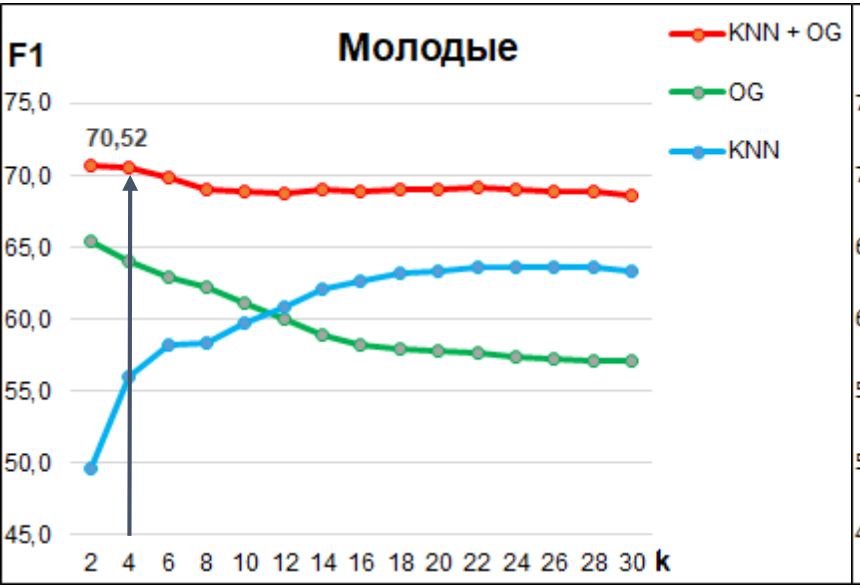
Метод предсказания функции на основе ортологии (OG) и k ближайших гомологов (KNN)

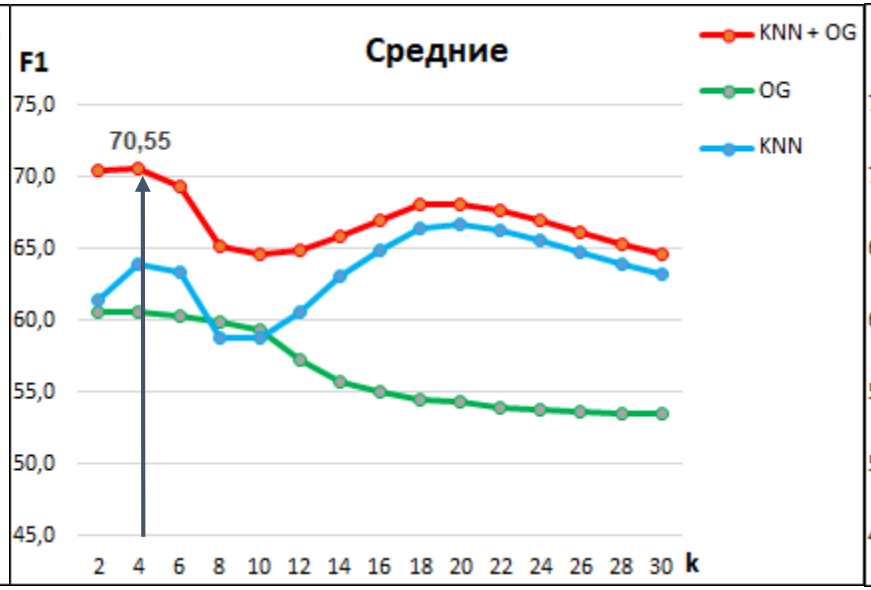


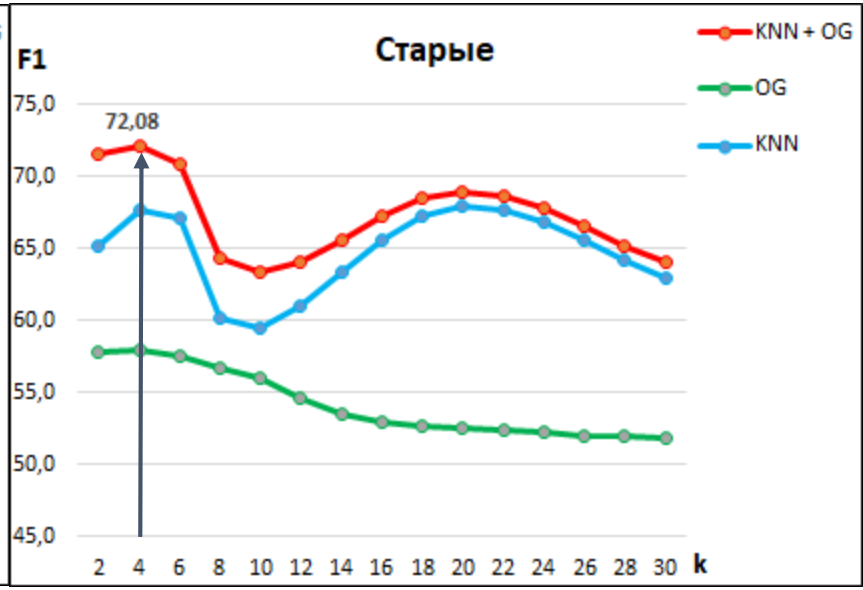
Этот метод опирается на концепцию ортологии. Для **семейств ортологов БД OrthoDB предоставляет аннотацию функции терминами GO**, которые можно использовать для предсказания функции искомого гена

Дополнительно мы использовали метод объединения предсказания функций KNN+OG

Точность предсказания функций для методов KNN, OG и KNN+OG







Объединение результатов метода 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=0,70-0,72)**

Сравнение с Blast2GO для *A.thaliana*

	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**

Δ=1,56

Δ=5,99

Δ=14,23

Таким образом на основании данных можно сделать вывод что в среднем **наш метод точнее Blast2GO более чем на 14%**

Время аннотации *A.thaliana*: Blast2GO - более 500 часов; KNN+OG = 2,5 часа

Точность для предсказания функций у других видов растений

Название вида	Молодые	Средние	Старые	Оптимальное значение k
Chlamydomonas reinhardtii	54.2	84.2	86.7	2
Oryza sativa	35.5	60.3	68.2	4
Solanum lycopersicum	59.7	73.4	76.4	6
Solanum tuberosum	57.7	73.1	74.1	6
Zea mays	38.6	56.1	58.6	4

Так для разных видов хорошо заметны **различия в F1 для разных возрастов**: средние и старые гены имеют высокие и близкие между собой значения, а молодые существенно меньше

При значении k = 4, точность близка к оптимальной практически для всех видов. Таким образом это значение было выбрано как универсальное

Выводы

1. Разработан метод KNN для анализа и предсказания функций генов растений по гомологии на основе *k* ближайших гомологов из базы данных OrthoDB, проведена оценка его точности на основе аннотации генов *Arabidopsis thaliana*, показано что оптимальные значения точности достигаются при идентичности последовательностей не менее 30% и покрытии не менее 60% и при числе ближайших соседей $k = 20$.
2. Показано, что возраст гена влияет на точность предсказания функции: для молодых генов лучшее значение F1 составило 63.68, для генов среднего возраста 66.58, для старых 67.53.
3. Предложен метод OG предсказания функций генов с учетом информации о принадлежности гена к ортологической группе OrthoDB, который при объединении с результатами KNN дает более высокую точность предсказания и позволяет нивелировать эффект возраста генов для *A.thaliana*: для молодых, средних и старых генов значения F1 составили 70.52, 70.55 и 72.08.
4. Сравнение предложенного метода с методом Blast2GO показало, что точность нашего подхода превосходит точность Blast2GO на величину от 18% и ниже.
5. Оценка точности метода для генов из 5 дополнительных видов показала, что для предсказания различных функций можно использовать одинаковые параметры алгоритма, однако результаты демонстрируют, что точность предсказания функций молодых генов оказывается ниже, чем для средних и старых.

СПАСИБО ЗА ВНИМАНИЕ