ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ «ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ИНСТИТУТ ЦИТОЛОГИИ И ГЕНЕТИКИ СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК»

На правах рукописи

КОМЫШЕВ ЕВГЕНИЙ ГЕННАДЬЕВИЧ

Разработка методов автоматического определения количественных характеристик, описывающих фенотипические признаки колоса пшеницы

Математическая биология, биоинформатика - 03.01.09

Диссертация на соискание учёной степени
кандидата биологических наук

Научный руководитель
к.б.н., доцент Д.А. Афонников

Новосибирск 2021

ОГЛАВЛЕНИЕ

Словарь сокращений 6

Введение 7

Актуальность работы 7

Цели и задачи исследования 9

Научная новизна работы 10

Теоретическая и практическая ценность работы 10

Положения, выносимые на защиту 11

Апробация работы 12

Объем и структура диссертации 13

Публикации 13

Благодарности 15

Глава 1. Обзор литературы 16

1.1 Объект исследования 16

1.1.1 Геном пшеницы 16

1.1.2 Структура растения пшеницы 19

1.1.3 Структура колоса пшеницы, характерные черты 21

1.1.4 Гены, определяющие морфологию колоса 25

1.1.5 Признаки пшеницы, связанные с урожайностью 29

1.2 Методы обработки биологических цифровых изображений 33

1.2.1 Основные процедуры подготовки и анализа изображений в биологии. 33

1.2.2 Сегментация, морфологические преобразования и поиск контуров 36

1.2.3 Задачи оптимизации 40

1.2.4 Задача распознавания образов 41

1.2.5 Применение программных средств для анализа данных фенотипа растений 43

1.3 Библиотеки анализа цифровых изображений 44

1.3.1 Matlab 45

1.3.2 Среда R для анализа изображений 46

1.3.3 Библиотека OpenCV 47

1.3.4 ImageJ 47

1.3.5 Scikit-image 48

1.4 Методы фенотипирования колосьев и зерен пшеницы 48

1.4.1 Фенотипирование колосьев 48

1.4.2 Анализ формы зерен 49

1.4.3 Морфометрия растений при помощи мобильных устройств 50

1.5 Базы данных и онтологии в области феномики, селекции и генетики растений 51

1.5.1 Феномика 52

1.5.2 Онтологии и их применение для решения задач биоинформатики 53

1.5.3 Gene Ontology 55

1.5.4 PlantOntology 56

1.5.5 Crop ontology 57

1.6 Статистический анализ данных 57

1.7 Заключение по обзору литературы и формулировка задачи исследования 59

Глава 2. Материалы и методы 61

2.1 Растительный материал 61

2.1.1 Растительный материал для морфометрии зерен 61

2.1.2 Растительный материал для морфометрии колосьев 61

2.2 Методы анализа изображений 63
2.3 Методы оценки точности алгоритмов анализа изображений 66
2.4 Методы статистического анализа 68
2.5 Разработка приложения для Android 69
2.6 Методы реализации баз данных 69
Глава 3. Результаты и обсуждение 71
3.1 Метод морфометрии зерен пшеницы с помощью мобильных устройств 71
3.1.1 Протокол получения изображения зерен пшеницы 71
3.2 Алгоритм анализа изображений для определения характерных черт зерен пшеницы 72
3.2.1 Распознавание листа бумаги 73
3.2.2 Идентификация и морфометрия зерен 74
3.3 Мобильное приложение SeedCounter 75
3.3.1 Интерфейс мобильного приложения 76
3.4 Оценка точности SeedCounter 77
3.5 Заключение по главе 3 87
Глава 4. Метод морфометрии колоса пшеницы 88
4.1 Протоколы получения изображений 88
4.2 Идентификация колоса и остей на изображении 89
4.2.1 Предварительная обработка изображения 89
4.2.2 Распознавание цветовой шкалы 89
4.2.3 Сегментация 90
4.2.4 Идентификация остей 91
4.2.5 Выбор параметров для выделения областей колоса и остей на изображении 93

4.2.6 Идентификация контура колоса и его выпрямление 94

4.2.7 Интегральные характеристики формы 96

4.3 Модель четырехугольников 97

4.4 Оценка точности распознавания областей остей и колоса 100

4.4.1 Анализ параметров остистости для выборки колосьев 106

4.5 Анализ характеристик формы колоса 107

4.5.1 Анализ корреляций между характеристиками 113

4.5.2 Анализ вариабельности морфометрических характеристик колоса 114

4.5.3 Вывод 120

4.6 Заключение по главе 4 120

Глава 5. Информационная система для аннотации морфометрических характеристик колоса пшеницы 123

5.1 Модель данных 123

5.2 Технологии реализации системы SpikeDroid 125

5.3 Модуль интерфейса системы SpikeDroid 126

5.4 Информационное содержание базы данных SpikeDroidDB 129

5.5 Заключение по главе 5 130

Заключение 132

Выводы 134

Список литературы 135

СЛОВАРЬ СОКРАЩЕНИЙ

ChIP-seq (ChIP-sequencing) - метод анализа ДНК-белковых взаимодействий, основанный на иммунопреципитации хроматина (ChIP)

CMF (Content Management Framework) - каркас веб-приложений

CMS (Content Management System) - система управления содержимым

EST (Expressed Sequence Tag) – экспрессирующий тег последовательности

EV (exposure value) - значение экспозиции

GWAS (Genome-Wide Association Study) - полногеномный поиск ассоциаций

HSV (Hue Saturation Value) - цветовая модель тон, насыщенность, яркость

QTL (Quantitative Trait Loci) - локусы количественных признаков

RGB (Red Green Blue) - цветовая модель красный, зеленый, синий

RNA-seq (RNA sequencing) - секвенирование РНК

SNP (Single nucleotide polymorphism) - однонуклеотидный полиморфизм

SNV (single nucleotide variants) - однонуклеотидный полиморфизм

БД – база данных

ОС – операционная система

СУБД – система управления базами данных

ВВЕДЕНИЕ

Актуальность работы

Одной из современных глобальных проблем человечества является рост населения и связанная с этим необходимость в увеличении производства продуктов питания. Согласно прогнозу Продовольственной и сельскохозяйственной организации ООН (FAO, Food and Agriculture Organization), производство зерновых культур должно удвоиться до 2050 года, чтобы удовлетворить спрос на продовольствие со стороны растущего населения мира. Следует также отметить, что в потреблении продуктов растениеводства возрастает конкуренция со стороны биотехнологических компаний, использующих их в промышленных целях как источников биоэнергии, волокна, крахмала и т.п. (Rahaman et al., 2015). Однако исследования климата показали, что наблюдаемые и прогнозируемые климатические изменения будут препятствовать повышению урожаев сельскохозяйственных культур за счет увеличения площади районов подверженных засухам и повышенным температурам (Sticklen, 2007). При этом расширение посевных площадей возможно в основном за счет районов, где условия возделывания менее пригодны для множества современных культур (Long and Ort, 2010). В связи с этим возрастает необходимость поиска новых, менее прихотливых к условиям выращивания сортов с более высокой продуктивностью и качеством урожая.

На решение описанных выше задач направлены современные методы генетики и селекции, основой которых является выявление связи между генотипом, окружающей средой и фенотипом. Благодаря быстрой разработке технологий секвенирования в молекулярно-генетических базах данных стали доступны целые геномы многих видов растений, основных сельскохозяйственных культур (Varshney et al., 2009), включая и мягкую пшеницу (Alaux et al., 2016). Технологии секвенирования позволяют относительно недорого и быстро осуществить определение геномных вариаций для тысяч отдельных растений. Из-за значительного дисбаланса накопленного объема генетических данных и недостатка

фенотипических, актуальным на данный момент является совершенствование технологий идентификации фенотипов растений.

Традиционные подходы к фенотипированию (оценка качественных признаков растений экспертами, измерения линейкой, взвешивание) точны и вполне приемлемы для многих возникающих задач, но являются трудоемкими и дорогостоящими, особенно когда в эксперименте проводится анализ сотен и тысяч растений. В таких условиях применение традиционных подходов к фенотипированию зачастую становится невозможным, что вынуждает исследователей ограничиваться меньшим масштабом эксперимента. Целью новых подходов к фенотипированию растений является повышение точности, производительности, сокращение трудозатрат и исключение человеческого субъективизма при проведении измерений за счет автоматизации и механизации (Hancock, 2014). Внедрение новых технологий сбора данных о фенотипических признаках растений позволит улучшить унификацию и интеграцию данных, полученных в результате экспериментов разными группами, в разное время и разных местах. Эти новые технологии основаны, прежде всего, на методах анализа цифровых изображений (Li et al., 2014).

В дополнение к разработке методов фенотипирования для хранения полученных данных возникает необходимость в стандартизации описания фенотипа растений в базах данных. В этой связи интенсивное развитие получили системы онтологий растений (Shrestha et al., 2012). Онтологии позволяют интегрировать различные подходы, методы, технологии и протоколы, которые могут быть задействованы для получения, обработки, хранения и анализа данных на всех этапах селекционно-генетических исследований.

Одной из важнейших сельскохозяйственных культур является мягкая пшеница (*Triticum aestivum* L.). На нее приходится более одной четвертой всего мирового производства зерновых культур, а также, она является главным источником основных продуктов питания для более чем одной пятой населения земного шара (Manske G. G. B. et al., 2001; FAO, 2011). Она также обеспечивает более 20 % калорий и белка для населения мира (Braun et al., 2010). Создание новых

высокопродуктивных сортов и линий пшеницы, устойчивых к биотическим и абиотическим стрессам позволит во многом обеспечить продовольственную безопасность существенной части населения земного шара.

Одними из важных признаков сельскохозяйственных растений являются признаки продуктивности. У мягкой пшеницы они связаны с размером и формой колоса, количества зерен и их массой (Farooq et al., 2015). Именно эти признаки в конечном итоге определяют урожайность растения. Однако высокопроизводительные методики для определения характеристик колосьев и зерен пшеницы в настоящее время недостаточно развиты. Это приводит к низкой производительности труда селекционера-генетика и затрудняет создание новых высокопродуктивных сортов и линий растений.

В этой связи актуальной задачей для современной генетики пшеницы является создание методов высокопроизводительного фенотипирования, которые позволили бы быстро и точно осуществлять оценку характеристик колосьев и зерен. Для эффективного хранения результатов фенотипирования необходимо развитие баз данных, в которых бы хранилась разнородная информация, включающая экспертные оценки фенотипа, изображения колосьев, количественные оценки фенотипа, полученные на основе анализа изображений.

Цели и задачи исследования

Целью работы является разработка методов автоматического определения количественных морфометрических характеристик колосьев и зерен пшеницы на основе анализа их цифровых изображений.

Для достижения заявленной цели были поставлены следующие задачи:

1. Разработка метода морфометрии зерен пшеницы с использованием мобильных устройств.
2. Разработка методов автоматического определения количественных характеристик формы и размера колоса на основе двухмерных изображений и его апробация на примере анализа колосьев пяти видов гексаплоидных пшениц.

3. Разработка базы данных для накопления, хранения и систематизации информации о фенотипических признаках колоса пшеницы.

Научная новизна работы

Впервые создано мобильное приложение SeedCounter для устройств под управлением ОС Android для подсчета зерен и определения их размеров.

Впервые предложен новый метод автоматического определения количественных морфометрических характеристик колоса пшеницы на основе анализа цифровых двухмерных изображений, который позволяет описывать форму колоса на основе модели четырехугольников.

Проведена оценка сходства и различий формы колосьев с использованием оценок, полученных у 14 образцов растений пяти генотипов мягкой пшеницы и их сородичей на основе параметров, оцененных путем анализа двухмерных цифровых изображений.

Разработана компьютерная система SpikeDroid, реализующая технологии хранения разнородной информации, для поддержки работы селекционера по сбору данных по морфометрическим характеристикам колосьев пшеницы и их диких сородичей.

Теоретическая и практическая ценность работы

Предложенные компьютерные методы позволяют на основании анализа цифровых изображений оценивать такие характеристики колоса пшеницы, как длина, ширина, остистость, плотность и тип колоса, количество зерен в колосе. Для зерен пшеницы оцениваются длина, ширина, проецируемая на поверхность площадь и ряд других характеристик формы и размера. Предложенные методы позволяют оценивать количественные характеристики продуктивности растений с высокой степенью детализации. Это позволяет исключить субъективизм присущий человеку при проведении измерений и не требует от пользователя специфических знаний в области фенотипирования пшеницы.

Мобильное приложение SeedCounter позволяет выполнять измерения в полевых условиях, без использования дополнительных технических средств, сохранять на

мобильном устройстве и отправлять данные на сервер посредством сети Интернет, с последующим экспортом данных в XML.

Разработанная система SpikeDroid позволяет существенно ускорить процесс массового фенотипирования благодаря автоматизации этапов начиная от получения изображений, заканчивая статистическим анализом занесенных в базу данных параметров.

Собранные воедино в системе SpikeDroid данные с анализируемых изображений, ручного фенотипирования и генотипов позволяют повысить эффективность существующих методов сравнительной генетики пшениц, таких как гибридологический метод анализа, метод возвратных скрещиваний, метод циклических скрещиваний и др.

Все эти характеристики разработанных методов позволяют существенно повысить эффективность селекционно-генетических экспериментов в направлении создания новых высокопродуктивных сортов и линий пшеницы.

Положения, выносимые на защиту

Методы фенотипирования колосьев и зерен пшеницы на основе анализа цифровых двухмерных изображений, реализованные в виде приложений WERecognizer и SeedCounter, позволяют проводить оценку характеристик продуктивности растений в автоматизированном режиме в массовых селекционно-генетических экспериментах.

Геометрическая модель колоса, описывающая его форму и размер в виде четырехугольников, предсказывает, что такие характеристики колоса, как длина, размер центральной части, ширина основания, площадь центрального сегмента и основания колоса являются наиболее значимыми для определения вида гексаплоидной пшеницы.

Компьютерная система SpikeDroid обеспечивает накопление, хранение, систематизацию и поиск информации о фенотипических признаках колоса, полученных из различных источников, а также доступ к ним через Web-интерфейс в сети Интернет.

Апробация работы

Работа представлена в виде устных и стендовых докладов на научных конференциях:

Komyshev E.G., Genaev M.A., Afonnikov D.A. SeedCounter – mobile and desktop application for high-throughput phenotyping seeds in wheat, BGRS\SB'2014 The 9th International conference on Bioinformatics of genome regulation and structure\System biology, Novosibirsk, Russia, June 23-28, 2014.

Komyshev E.G., Genaev M.A., Afonnikov D.A. SeedCounter – mobile application for grain phenotyping // 3-я Международная конференция "Генетика, геномика, биоинформатика и биотехнология растений", PlantGen, Новосибирск, 2015.

Комышев Е.Г., Генаев М.А., Афонников Д.А. SeedCounter – мобильное и настольное приложение для массового фенотипирования зерен пшеницы // Актуальные проблемы вычислительной и прикладной математики 2015, AMCA, 2015.

Komyshev E.G., Genaev M.A., Afonnikov D.A. SeedCounter application and wheat ear recognizing algorithm for high throughput wheat phenotyping // Международный научный симпозиум «Генетика и геномика растений для продовольственной безопасности» - Institute of Cytology and Genetics SB RAS, 26-28 August 2016, Novosibirsk, Russia.

Komyshev E.G., Genaev M.A., Akushkina A.V., Afonnikov D.A. Wheatdb2: plant trait database and information system based on CropOntology terms // BGRS\SB'2016 The 10th International conference on Bioinformatics of genome regulation and structure\System biology, Novosibirsk, Russia, 29 August - 2 September, 2016.

Комышев Е.Г., Генаев М.А., Акушкина А.В., Афонников Д.А. Приложение SeedCounter и алгоритм распознавания колоса пшеницы для высокопроизводительного фенотипирования // Всероссийская Конференция «50 лет ВОГиС: успехи и перспективы», 8-10 ноября 2016 г., Москва.

Komyshev E., Genaev M., Tumanyan S., Goncharov N., Afonnikov D., Koval V. Wheat ear recognizing algorithm for high throughput wheat phenotyping BGRS\SB'2018,

The 11th International Conference On Bioinformatics of Genome Regulation and Structure\Systems Biology, 20-25 August, 2018.

Комышев Е.Г., Генаев М.А., Афонников Д.А. Метод морфометрии колоса пшеницы на основе анализа изображений \\ Международный конгресс биотехнология: состояние и перспективы развития, 25 - 27 февраля 2019, Ильинка, 4, Гостиный двор, Москва.

Komyshev E.G., Genaev M.A., Smirnov N.V., Afonnikov D.A. Cereal signs analysis associated with color on digital images // 5th International scientific conference Plant genetics, genomics, bioinformatics and biotechnology (PlantGen2019) June 24-29, 2019, Novosibirsk, Russia.

Komyshev E.G., Genaev M.A., Afonnikov D.A., Kruchinina Y.V., Koval V.S., Goncharov N.P. Spikes Morphometric Characteristics Analysis of Five Species of Wheat // BGRS/SB-2020: 11th International Multiconference “Bioinformatics of Genome Regulation and Structure/Systems Biology“, 6-10 July 2020, Novosibirsk, Russia

Комышев Е.Г., Генаев М.А., Афонников Д.А. Фенотипирование колосьев пшеницы на основе анализа цифровых изображений // Международная конференция «Марчуковские научные чтения 2020» (МНЧ-2020), посвященная 95-летию со дня рождения академика Гурия Ивановича Марчука. Академгородок, 19 - 23 октября 2020 г., Новосибирск, Россия.

Комышев Е.Г., Генаев М.А., Афонников Д.А. Анализ морфометрических характеристик колосьев пяти видов пшеницы // 5-я международная конференция “Генофонд и селекция растений” 11-13 ноября 2020 г., Новосибирск, Россия

Объем и структура диссертации

Работа состоит из введения, обзора литературы, глав, посвященных используемым материалам и методам, разработанным новым методам, и результатам, а также из заключения, выводов и списка литературы. Работа содержит 36 рисунков и 17 таблиц. Список литературы содержит 135 источников.

Публикации

По материалам диссертации опубликовано 17 работ, из них 5 статей в рецензируемых научных журналах, входящих в список ВАК:

Афонников Д.А., Генаев М.А., Дорошков А.В., **Комышев Е.Г.**, Пшеничникова Т.А. Методы высокопроизводительного фенотипирования растений для массовых селекционно-генетических экспериментов //Генетика. – 2016. – Т. 52. – №. 7. – С. 788-803.

Komyshev E.G., Genaev M.A., Afonnikov D.A. Evaluation of the SeedCounter, a mobile application for grain phenotyping //Frontiers in plant science. – 2017. – Т. 7. – С. 1990.

Генаев М.А., **Комышев Е.Г.**, Фу Хао, Коваль В.С., Гончаров Н.П., Афонников Д.А. SpikeDroidDB – информационная система для аннотации морфометрических характеристик колоса пшеницы // Вавиловский журнал генетики и селекции. - 2018. 22(1):132-140. DOI 10.18699/VJ18.340

Genaev M.A., **Komyshev E.G.**, Smirnov N.V., Kruchinina Y.V., Goncharov N.P., Afonnikov D.A. Morphometry of the Wheat Spike by Analyzing 2D Images // Agronomy. – 2019.

Пронозин А.Ю., Паулиш А.А., Заварзин Е.А., Приходько А.Ю., Прохошин Н.М., Кручинина Ю.В., Гончаров Н.П., **Комышев Е.Г.**, Генаев М.А. Автоматическое фенотипирование морфологии колоса тетра- и гексаплоидных видов пшеницы методами компьютерного зрения // Вавиловский журнал генетики и селекции. – 2021. – Т. 25. – №. 1. – С. 71-81.

Получено два авторских свидетельства:

Свидетельство о государственной регистрации программы для ЭВМ №2014662191, «Мобильное и настольное приложение для массового фенотипирования зерен пшеницы (СиидКаунтер) / Mobile and desktop application for mass phenotyping seeds of wheat (SeedCounter)».

Свидетельство о государственной регистрации программы для ЭВМ №2019666362, «Программа для оценки количественных характеристик колоса

пшеницы (ОКХК) / The program for the extraction the quantitative characteristics of the wheat spike (WERecognizer)».

Личный вклад соискателя

Основная часть работы выполнена автором самостоятельно. В работах по созданию метода фенотипирования колосьев пшеницы автор принимал участие в разработке алгоритма анализа изображений, ручной разметки изображений, статистическом анализе полученных данных, оценке точности.

В работах по созданию метода фенотипирования зерен пшеницы автор разрабатывал мобильное приложение SeedCounter, проводил оценку точности, проводил статистический анализ полученных данных.

Автор принимал участие в создании модели, описывающих форму колоса пшеницы, проведении вычислительных экспериментов, обсуждении и анализе полученных результатов.

Благодарности

Автор выражает искреннюю благодарность научному руководителю работы к.б.н. Афонникову Д.А., соавторам и коллегам по работе – к.б.н. Генаеву М.А., к.б.н. Ковалю В. С., к.б.н. Дорошкову А.В., к.б.н. Пшеничниковой Т.А., д.б.н. Гончарову Н.П. за предоставление биологического материала, консультации, ценные советы и помощь в биологической интерпретации результатов. Автор выражает благодарность Николаю Смирнову за алгоритм распознавания цветовой шкалы ColorChecker и алгоритм цветокоррекции, использованные в методах морфометрии колосьев и Фу Хао за экспертную оценку фенотипических характеристик колосьев пшеницы. Коваль В.С. (протокол съемки колосьев, получение изображений), Кручинина Ю.В. (массовое фенотипирование колосьев).

Часть работ была выполнена при финансовой поддержке РФФИ (мол_а 14-07-31226, мол_а 16-37-00304), РНФ (грант №17-74-10148) и Курчатовского геномного центра Федерального исследовательского центра ИЦиГ СО РАН, соглашение с Министерством образования и науки РФ № 075-15-2019-1662.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

1.1 Объект исследования

Пшеница (род *Triticum*) является одной из наиболее значимых сельскохозяйственных культур, важной компонентой рациона питания миллиардов человек по всему миру. Она богата ценными белками, клетчаткой, различными ферментами, а также витаминами и микроэлементами. Общая площадь посевов пшеницы во всем мире - 210 миллионов гектаров (Zeybek and Yigit, 2004).

В связи с проблемой изменяющегося климата и роста населения Земли актуальным является создание новых сортов и линий пшеницы устойчивых к болезням, высокоурожайных и имеющих высокое качество зерна. Несмотря на то, что современные сорта пшеницы довольно устойчивы к заморозкам и другим абиотическим стрессам, они подвержены таким заболеваниям как бурая и желтая ржавчина, пыльная головня, мучнистая роса и др. (Гультяева и др.; Санин и др.). Заболевания растений приводят к существенным потерям урожая, которые могут достигать миллионов тонн зерна в год (Karakas and Gurel, 2010). Поэтому важным направлением остается создание сортов устойчивых как к биотическим, так и абиотическим факторам.

Эти задачи могут быть эффективно решены на основе использования современных достижений генетики, внедрения технологий массового фенотипирования и использования современных информационных технологий.

1.1.1 Геном пшеницы

Генетический потенциал рода *Triticum* огромен. Род включает в себя ди-(2n=14), тетра-(2n=28) и гексаплоидные (2n=42) виды. Из них более двадцати- это “естественные” виды (произрастающие в природе либо возделываемые человеком), и около десяти видов – искусственно полученные амфиплоиды (Гончаров, 2012). Ни один злак не имеет столько видов и сортов, как пшеница (таблица 1).

Полиплоидия - важная эволюционная черта, широко распространенная в царстве растений, возникающая в результате дупликации генома, после чего отдельные гены эволюционируют независимо. Аллополиплоидизация (рисунок 1) пшеницы привела

к появлению видов с лучшими агрономическими показателями и широкой адаптивностью (Chalupska et al., 2008).

Таблица 1. Классификация рода *Triticum* (Гончаров, 2009)

Секция	**Группа**	**Вид**	**2n**	**Геном**
Monococcon Dum.	Пленчатые	*T. urartu* Thum. Ex Gandil.	14	A^u
		T. boeoticum Boiss.	14	A^b
		T. monococcum L.	14	A^b
	Голозерная	*T. sinskajae* A.Filat. et Kurk.	14	A^b
Dicoccoides Flaksb.	Полбы	*T. dicoccoides* (Korn. Ex Aschers. Et Graebn.) Schweinf.	28	BA^u
		T. dicoccum (Schrank) Schubl.	28	BA^u
		T. karamyschevii Nevski	28	BA^u
		T. ispahanicum Heslot	28	BA^u
	Голозерные	*T. turgidum* L.	28	BA^u
		T. durum Desf.	28	BA^u
		T. turanicum Jakubz.	28	BA^u
		T. polonicum L.	28	BA^u
		T. aethiopicum Jakubz.	28	BA^u
		T. carthlicum Nevski	28	BA^u
Triticum	Пленчатые	*T. macha* Dekapr. et Menabde	42	BA^uD
		T. spelta L.	42	BA^uD
		ssp. tibetanum (Shao) N. Gontsch.		
		ssp. Yunnanense (King) N. Gontsch.		
		T. vavilovii (Thum.) Jakubz.	42	BA^uD
	Голозерные	*T. compactum* Host	42	BA^uD
		T. aestivum L.	42	BA^uD
		ssp. aestivum		
		ssp. hadropyrum (Flaksb.) Tzvel.		
		ssp. petropavlovskyi (Udacz. et Migusch.) N. Gontsch.		
		T. sphaerococcum Perciv.	42	BA^uD
Timopheevii A.Filat. et Dorof.	Пленчатые	*T. araraticum* Jakubz.	28	GA^u
		T. timopheevii Zhuk.	28	GA^u
		ssp. militinae (Zhuk. et Migusch.) N. Gontsch.		
		T. zhukovskyi Menabde et Erizjan	42	GA^uA^b
Compositum N. Gontsch.	Aegilotricum	*T. palmovae G. Ivanov (sun. T. erebuni Gandil.)*	28	DA^b
	Пленчатые	*T. dimococcum Schieman et Staudt*	42	BA^uA^b
		T. soveticum Zhebrak	56	BA^uGA^t
		T. kiharae Dorof. Et Mihusch.	42	GA^uD
		T. borisovii Zhebrak	70	BA^uDGA^t
	Голозерный	*T.flaksbereri Navr.*	56	GA^tBA^u

Вся пшеница принадлежит к роду *Triticum*, члену семейства злаковых (Gramineae или Poaceae). Ячмень (Hordeum vulgare L.) и рожь (Secale cereale L.) принадлежат к одному и тому же племени Hordeae, у которого один или несколько цветковых колосков сидячие и чередуются на противоположных сторонах рахиса (главной оси соцветия), образуя колос. Они также являются близкими

родственниками некоторых сорняков, таких как Agropyron и других диких трав, которые можно скрещивать с пшеницей (Thinopyrum, Leymus, Aegilops). Эту родственную группу злаковых часто называют Triticeae, что определяется ее родством с пшеницей. Различные виды Triticeae адаптированы к широкому диапазону условий внешней среды: от степных и полузасушливых, до горных и влажных регионов.

Диплоидные виды Triticeae имеют общее количество (семь пар) хромосом, унаследованных от общего предка. Таким образом, даже если происходят эволюционные процессы, такие как транслокации (изменение порядка генов или содержания генов), производные гомеологические хромосомы по-прежнему имеют большое сходство между различными видами Triticeae (Konopatskaia et al., 2016).

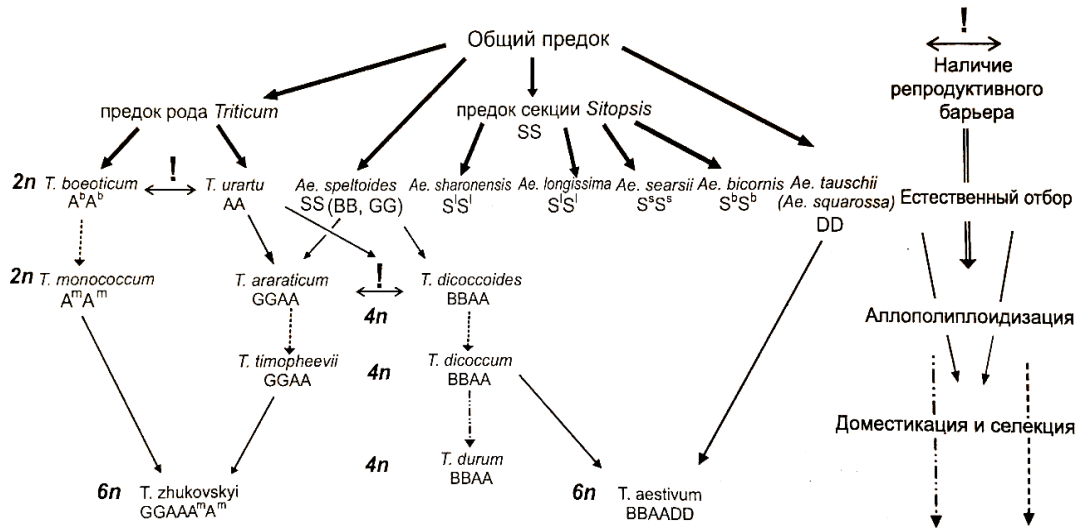


Рисунок 1. Вероятная схема происхождения филогенетических линий Emmer и Timopheevii пшениц (Гончаров, 2009)

Недавно были получены геномные последовательности нескольких видов рода *Triticum*, включая *Triticum urartu* (Ling et al., 2013), *Triticum aestivum* (International Wheat Genome Sequencing Consortium et al., 2014; Appels et al., 2018), *Aegilops tauschii* (Jia et al., 2013). Аллогексаплоидный геном пшеницы (2n = 6x = 42) является одним из самых крупных среди выращиваемых культур с размером гаплоидного набора в 16 миллиардов оснований, из которых повторы составляют около 80% (Bennett and Leitch, 1997). Сложная, аллогексаплоидная природа генома мягкой пшеницы, с

одной стороны, затрудняет генетический анализ, с другой стороны, позволяет применять уникальные подходы сравнительной генетики, использующие свойства гомеологичных хромосом (Goncharov, 2002). Изучение столь сложного генома пшеницы требует комплексного подхода.

1.1.2 Структура растения пшеницы

Продуктивность растения – комплексный признак, который зависит от многих факторов, структуры и физиологических характеристик растения.

Современные высокопродуктивные сорта пшеницы имеют более мощный фотосинтетический аппарат, ассимиляционный потенциал которого обусловлен большими площадью и удельной массой листьев, повышенным содержанием в них хлорофилла, длительностью активного функционирования листовой поверхности в течение вегетации, по сравнению со стародавними сортами и их дикорастущими предшественниками. Дальнейший прогресс в увеличении общей биологической продуктивности пшеницы связан с улучшением количественных характеристик колоса, а также других частей растений (стебля и листьев), обеспечивающих высокий урожай для тех или иных условий произрастания.

Характеристики колоса, такие как число колосьев на растение, число колосков в колосе, количество зерен в колосе, абсолютный вес зерна, обмолачиваемость колоса и др. являются важнейшими компонентами селекции на продуктивность у мягкой пшеницы. Также важны биомасса колоса, поэтому наряду с селекцией сортов с оптимальной длиной стебля, отдельных междоузлий, необходимо достижение определенной величины длины колоса, числа колосков и их озернённости. Помимо этого, среди факторов, связанных с продуктивностью, можно выделить длину стебля пшеницы. Это связано с затратами на транспорт метаболитов и дыхание, рост и поддержание жизнедеятельности (Степанов и др., 2008).

Рассмотрим подробней структуру растения пшеницы. Полностью развитое растение пшеницы состоит из главного стебля с колосом, междоузлий, узлов, листьев, корней и побегов. В свою очередь, каждый побег также состоит из колоса, междоузлий, узлов, листьев, корней и (потенциально) вторичных побегов. Узел – это

участок оси побега растений (стебля), на котором образуются боковые органы. Соответственно междоузлия - это участки между двумя смежными узлами.

У хлебных злаков нет главного стержневого корня. При прорастании зерна появляется несколько придаточных корней (Флякcбергер, 1922). Стебель представляет собой более или менее цилиндрическую соломину с узлами и полыми междоузлиями (рисунок 2).

Листья простые, линейные, двурядные, очередные, каждый из них отходит от узла. Лист состоит из 2-х основных частей: пластинки (самого листа) и влагалища. Влагалище – представляет собой свернутую вокруг стебля пластинку, не сросшуюся своими краями и утолщенную у основания.

На сгибе листа со стороны, которая прилегает к стеблю, может быть тонкий полупрозрачный придаток – язычок. У некоторых злаков на месте сгиба листа отростки – ушки, особенно развитые у ячменей. На границе между пластинкой и влагалищем располагаются три выроста: пленчатый язычок, который прилегает к стеблю, и пара охватывающих его пальцевидных ушек.

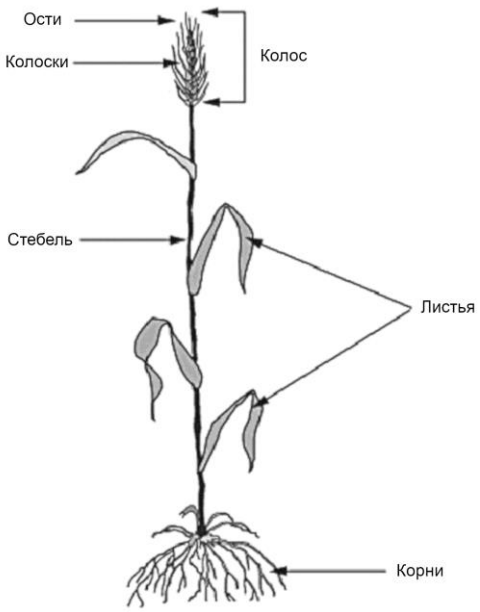


Рисунок 2. Схема структуры побега пшеницы. В нижней части рисунка изображены придаточные корни растения. От них над грунтом возвышается стебель из узлов и полых междоузлий. К узлам прикрепляются листья. В верхней части растения изображен сложный колос, состоящий из стержня, колосков и ответвляющихся от них остей.

Урожайность пшеницы прежде всего зависит от характеристик репродуктивных органов – соцветия, колоса и зерен.

1.1.3 Структура колоса пшеницы, характерные черты

Верхнее междоузлие, которое еще называют цветоносом, несет соцветие – сложный колос. Колос является важнейшим органом растения, непосредственно связанным с таким хозяйственно важным выходным параметром как урожайность.

Колос пшеницы состоит из многоцветковых колосков. Стоит отметить, что такая структура встречается не у всех злаков. К примеру, колос ячменя содержит одноцветковые колоски. Он включает в себя членики стержня, образующие коленчатую центральную ось, и отходящие от нее простые соцветия – колоски, которые обращены к оси широкой стороной. Колоски своим основанием крепятся к членикам стержня. Колосок состоит из двух наружных колосковых чешуй: нижней (ость) и верхней. Колосковые чешуйки обхватывают или частично прикрывают один или несколько цветков с цветочными пленками (рисунок 3).

Рисунок 3. Структура колоса пшеницы. Колос с частью отрезанного стебля изображен на черном фоне. Отмечены структурные элементы: а) Членики стержня колоса. б) Колоски в) Верхушечный колосок г) Ости.

Колоски на стержне располагаются поочередно то влево, то вправо. Обычно на колосе с боковой стороны можно распознать два ряда колосков, так как шестирядные сорта пшеницы встречаются довольно редко. С лицевой же стороны колоски налегают один на другой, поэтому часто лицевая сторона называется черепитчатой (рисунок 4).

Рисунок 4. Колосок. А) Цветки. Б) Нижняя цветковая чешуя. В) Верхняя цветковая чешуя. Г) Ости.

Зерновки (плод, зерно) у пшеницы (также, как и у других представителей триба *Triticeae*, а также триба *Aveneae*) можно различить по наличию углубленной угловой области на вентральной стороне, простирающейся по всей длине зерна и имеющей наибольшее углубление в середине (Evers and Millar, 2002). Зерновка состоит из семени, к которому плотно прирастает тонкий околоплодник. Семя состоит из семенной кожуры, большого мучнистого белка, состоящего из крупных клеточек, наполненных зернами крахмала, и небольшого зародыша, лежащего внизу почти под самой семенной кожурой в углублении белка (рисунок 5).

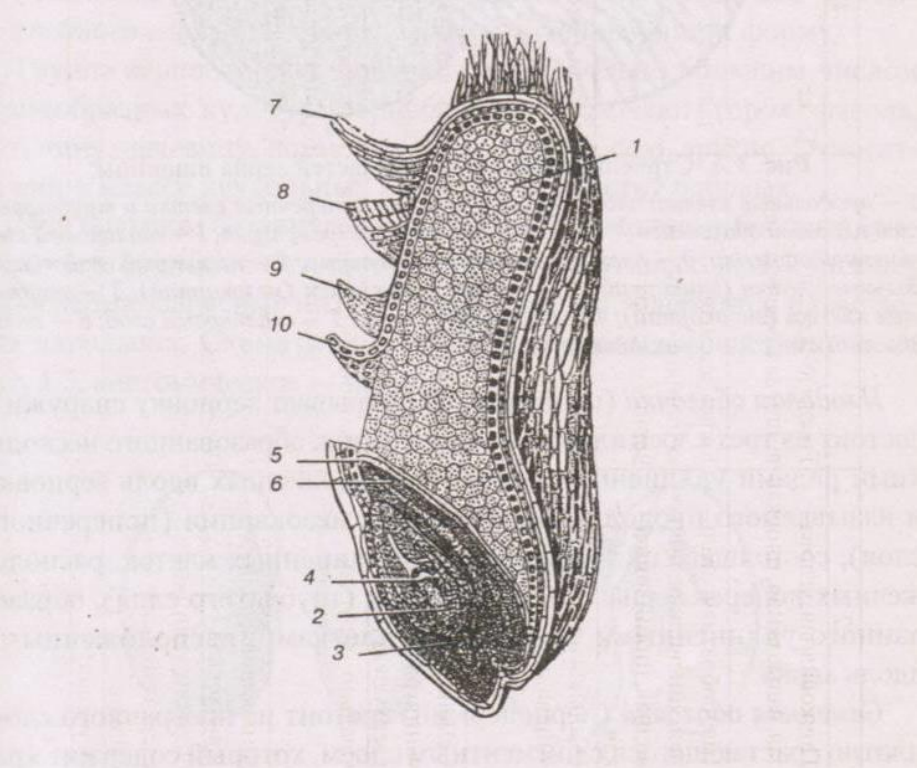


Рисунок 5. Строение зерна пшеницы – продольный разрез. 1 – эндосперм; 2 – зародыш; 3 – корешок; 4 – зачаточный лист; 5 – щиток; 6 – цилиндрический эпителий; 7 и 8 – плодовые оболочки; 9 – семенная оболочка; 10 – алейроновый слой (Кошкин и др., 2005).

Белок окружен клейковинным слоем клеточек (у ячменя в несколько слоев), содержащий зернистое белковое вещество. Зародыш состоит из почечки, семядоли, обращенной в сторону белка, щитка - чешуевидного отростка с наружной стороны у основания почечки (отсутствует у ржи и ячменя), и 3-8 корешков. С лицевой стороны можно распознать зерновки на каждом уровне колоса.

Существует различные классификации формы колоса. Например, ранняя классификация (Флякcбергер, 1922), основанная на геометрическом подобии форм:

- веретеновидные - средняя часть колоса наиболее широкая, к вершине и частично к основанию сужается;
- эллиптические – колосья формы вытянутого круга;
- призматические – колосья почти одинаковой ширины по всей длине, не считая областей верхнего и нижнего колосков;

- удлиненно конусовидные – колосья, суживающиеся к вершине от основания;
- булавовидные – расширяющиеся к вершине колосья;
- цилиндрические – колосья, имеющие одинаковый радиус по всей длине.

По длине колосья подразделяются: у мягкой пшеницы на мелкие (до 8 см длины), средние (8—10 см) и крупные (длиннее 10 см); у твердой пшеницы на короткие (до 6 см), средней длины (7—8 см), удлиненные (8-9 см), крупные (10 см и более).

На данный момент, в связи с изученностью генов, контролирующих характерные признаки колоса, наиболее актуальной является классификация на следующие формы: компактные – широкие колосья, небольшой длины, спельтоидные – удлинненные узкие колосья и нормальные – колосья средней длины и ширины (рисунок 6) (Гончаров, 2012).

Рисунок 6. Типы формы колоса. А) Компактный Б) Нормальный В) Спельтоидный

В данной работе, мы будет придерживаться именно этой классификации при определении типа формы колоса. Признаки, определяющие один из этих типов формы, как было показано (Swaminathan and Rao, 1961), контролируются небольшим числом генов, описанных далее.

1.1.4 Гены, определяющие морфологию колоса

Современные культивируемые пшеницы отличаются от своих диких родственников и предков набором морфологических и физиологических особенностей, включая черты, связанные с морфологией колоса, такие как: форма колоса, ломкоколосость и обмолачиеваемость (легкость обмолота), наличие или отсутствие остей, ветвление и наличие дополнительных колосков.

1.1.4.1 Легкость обмолота

Обмолачиваемость колосьев в основном зависит от прочности чешуек, покрывающих зерна. Дикие виды обладают зернами, покрытыми жесткой чешуей (non free-threshing), которая остается налипшей на зерно после обмолота (Dorofeev et al., 1979). Голые или свободно обмолачиваемые семена культурных сортов пшеницы окружены мягкой чешуей, которая выделяется при обмолоте. Сорта пшеницы со свободно обмолоченными семенами появились во время одомашнивания и значительно повысили эффективность процесса обмолота (Zhang et al., 2011).

Ранее были идентифицированы несколько основных генов и локусов, связанных с обмолачиваемостью: ген спельтоидности (*Q*), локус цепкой чешуи (*Tg*), локус цепкой чешуи 2 (*Tg2*) и локус мягкой чешуи (*sog*) (Konopatskaia et al., 2016).

Ген *Q* - *в*ажнейший ген, определяющий морфологию колоса. Ген *Q* (синонимы: WAP2, wheat AP2), расположенный на длинном плече хромосомы 5A, кодирует AP2-подобный фактор транскрипции и играет важную роль в одомашнивании полиплоидной пшеницы (Zhang et al., 2011). Ген *Q* регулирует обмолоченность полиплоидных видов пшеницы и оказывает плейотропное действие на несколько других важных характеристик, включая форму колоса, хрупкость стержня, высоту растения и время цветения (Konopatskaia et al., 2016). Полиплоидные виды пшеницы имеют дополнительные гомеологические локусы для гена *Q* на хромосомах 5B и 5D.

5AQ играет главную роль в придании признаков, связанных с одомашниванием; *5Dq* вносит прямой вклад, а *5Bq* косвенно - в подавление фенотипа спельтоидов (Zhang et al., 2011); *5Bq* представляет собой псевдоген, который не кодирует полноразмерный белок *q*, но участвует в регуляции экспрессии *5AQ* и *5Dq*; комбинация локусов *5AQ*, *5Bq* и *5Dq* важна для образования колосьев свободного обмолота (Simons et al., 2006).

Локус (*Tg*) был описан для гексаплоидных видов пшеницы, локализованный на коротком плече хромосомы 2D (Kerber and Rowland, 1974; Jantasuriyarat et al., 2004; Sood et al., 2009).

У тетраплоидных видов пшеницы дополнительный ген, который влияет на обмолот (*Tg2*), расположен на хромосоме 2BS (Simonetti et al., 1999). Таким образом, признак легкого обмолота у тетраплоидных видов пшеницы является сложным, и предполагается, что в его формировании участвуют *Q*, *Tg2* и несколько минорных генов.

1.1.4.2 Остистость

Ости - это нитевидные расширения нижней цветковой чешуи (леммы). Они выполняют фотосинтетическую функцию, увеличивают усвоение воды и могут способствовать высокому урожаю пшеницы, выращенной в условиях ограниченного количества воды. Остистость колоса подразделяется на безостые, полуостистые, коротко-остистые, длинно-остистые, в большинстве случаев лишь субъективно.

Фенотип безостых колосьев контролируется тремя неаллельными генами *Hd (Hooded), B1 (Tipped 1) и B2 (Tipped 2)*, локализованными на хромосомах 4AS, 5AL и 6BL, соответственно (Rao, 1981; Sears, 1966). Их различные комбинации вносят изменения в характеристики ости. У гексаплоидных видов пшеницы доминирующий ген *B1* играет основную роль в формировании безостых колосьев. Исключение составляют лишь несколько образцов *T. aestivum* из Китая и Индии, у которых доминантные гены *B2* и *Hd* определяют колосья с характерными типами остей (Гончаров, 2012).

B1 дает фенотип с апикальными остями на вершине и отсутствием у основания и середины колоса. Кончики остей *B1* обычно прямые и отогнутые у основания,

тогда как ости *B2* слегка изогнуты и почти равны по длине вдоль колоса. Для *Hd* ости уменьшены в длине, изогнуты/скручены и в некоторых случаях значительно расширены у основания, напоминая перепончатое латеральное расширение мутантов *Hooded* у ячменя (Wang et al., 2019).

Локус *B1* расположен на длинном плече хромосомы 5A, в области, содержащей только два предсказанных гена, включая фактор транскрипции цинкового пальца C2H2 (*AWNS-A1*) 219 п.н. из 5A28417 (DeWitt et al., 2020).

Ген ингибирования ости *B1* (*TraesCS5A02G542800*, ген цинкового пальца C2H2) был идентифицирован с использованием объемного сегрегантного РНК-секвенирования популяции F2 твердой пшеницы и путем делеционного картирования мутантов остистой мягкой пшеницы. *B1* был охарактеризован посредством избыточной экспрессии конститутивного гена как в твердой, так и в мягкой пшенице. Были приведены доказательства его роли в качестве репрессора транскрипции (Huang et al.).

Ген цинкового пальца C2H2, ингибитор длины *Awn Length Inhibitor 1 (ALI-1)*, был также независимо идентифицирован в работе Ван и др. (Wang et al., 2019). *ALI-1* транскрипционно подавляет гены, расположенные ниже по течению, снижает содержание цитокининов и одновременно сдерживает передачу сигнала, что приводит к уменьшению количества клеток в ости. Помимо этого, *ALI-1* регулирует развитие зерна, а также влияет на количество колосков в колосе (Wang et al., 2019).

1.1.4.3 Ломкоколосость

Признак ломкоколосости определяет то, насколько легко колосья расчленяются на отдельные колоски, что в свою очередь зависит от места разделения. Естественный распад зрелого колоса позволяет распределять семена пшеницы на большее расстояние. Помимо гена *Q*, три локуса *Br1, Br2 и Br3*, расположенные на хромосомах гомеологической группы 3, контролируют свойства стержня колоса (Konopatskaia et al., 2016). Локус *Br* локализован на коротком плече хромосомы 3D у тибетских староместных сортов мягкой пшеницы, и на коротком плече хромосом 3A у T. timopheevii (Li and Gill, 2006). Рецессивные аллели *Br2 и Br3*, расположенные на 3AS и 3BS, определяют нехрупкий стержень у тетраплоидов. Аллели *Br1, Br2* и

Br3 не были клонированы на молекулярном уровне, и ортологи этих генов в зерновых не известны.

1.1.4.4 Форма колоса

Колосья пшеницы можно отнести к одному из трех морфологических вариантов формы (Dorofeev et al., 1979):

• нормальные (колосья с параллельными сторонами и относительно короткой, квадратной верхушкой);

• спельта, полба (пирамидальные колосья с удлиненным стержнем и цепкими чешуйками);

• компактный (короткий, плотный колос с уменьшенным числом колосков).

В контроле формы колоса у видов пшеницы участвуют три основных гена/локуса: *Q, C, C2* (Konopatskaia et al., 2016). Аллели гена *5AQ* определяют нормальную и спельтоидную форму колоса у разных видов пшеницы. Гомологичные *5Dq* и *5Bq* также способствуют подавлению спельтоидного фенотипа (Zhang et al., 2011).

Локализация гена *C2* в геноме остается неизвестной. Признак компактоидной формы шипа у тетраплоидов контролируется двумя неаллельными рецессивными генами, названными sc1 и sc2 (Goncharov, 1997). Один только ген *sc1* отвечает за полукомпактоидный колос, в то время как вместе с геном *sc2* приводит к образованию компактоидного колоса у тетраплоидной пшеницы. Последовательности генов *C, C2, sc1* и *sc2* не были клонированы на молекулярном уровне.

1.1.4.5 Ветвление

Обычные колосья пшеницы состоит из колосков, расположенных в два противоположных ряда вдоль стержня. В ветвящихся колосьях, колоски заменяются боковыми ветвистыми структурами, напоминающими вторичные колосья небольшого размера. Образование ветвей дает значительно больше зерен на колос и, таким образом, увеличивает урожай зерна (Konopatskaia et al., 2016). Фенотип разветвленных колосьев распространен у тетраплоидных видов *T. turgidum*, и,

предположительно, этот признак контролирует главный локус bh^t, расположенный на коротком плече хромосомы 2A (Haque et al., 2011).

1.1.5 Признаки пшеницы, связанные с урожайностью

Для злаковых урожайность является важнейшей характеристикой растений. Существует несколько признаков, характеризующих урожайность растения: количество колосьев; компактность, вес, размер, форма и тип колоса; количество зерен в колосе; размеры и масса зерен; стекловидность и т.д. При оценке урожайности пшеницы оцениваются такие характеристики как число зерен в колосе, размер зерен, их плотность, масса 1000 зерен, количество плодородных колосков (колоски с ненулевым количеством зерновок). При оценке ряда признаков, таких как компактность и компактоидность колоса, подсчитывается плотность колоса, D (Фляксбергер К. А., 1935):

$$D = \frac{(\text{число колосков} - 1) * 10}{\text{длина оси колоса, см.}} \quad (1)$$

Помимо этого, урожайность, а также экономическая ценность определяется также такими факторами как расстояние между рядами растений (плотность взращивания) и комбинирование различных сортов при взращивании. Поэтому для объективной оценки ценности сортов исследователи используют широкий набор характеристик и показателей, наиболее полно отражающих продуктивность пшеницы.

В генетических исследованиях, в которых выявляют локусы количественных признаков (Quantitative Trait Loci, QTL), связанных с урожайностью пшеницы, определяются десятки признаков фенотипа для сотен (тысяч) образцов растений. В произведении это приводит к десяткам (сотням) тысяч измерений.

Высокая трудоемкость проведения таких измерений ограничивает размер популяции растений, задействованных в эксперименте. Рассмотрим примеры нескольких недавних исследований на эту тему.

В работе (Guo et al., 2017) проведен полно-геномный поиск ассоциации геномных вариантов и 54 фенотипических признаков (Genome-Wide Association Study, GWAS): 16 признаков фертильности цветков и 38 признаков морфологии

колосьев и ассимиляционного распределения, у 210 европейских образцов озимой пшеницы.

Результаты эксперимента указывают на потенциальные взаимосвязи, выявленные анализом QTL, между фертильностью цветков, ассимиляционным распределением и морфологией колосьев. Фенотипические характеристики колосьев могут обуславливать несколько генов-кандидатов, вовлеченных в метаболизм углеводов, фитогормонов или развитие цветков. Они находятся вблизи обнаруженных QTL. В работе предложена генетическая сеть, включающая такие гены и лежащая в основе фертильности цветков и связанных с ней признаков, что позволяет определять детерминанты для улучшения показателей урожайности (рисунок 7).

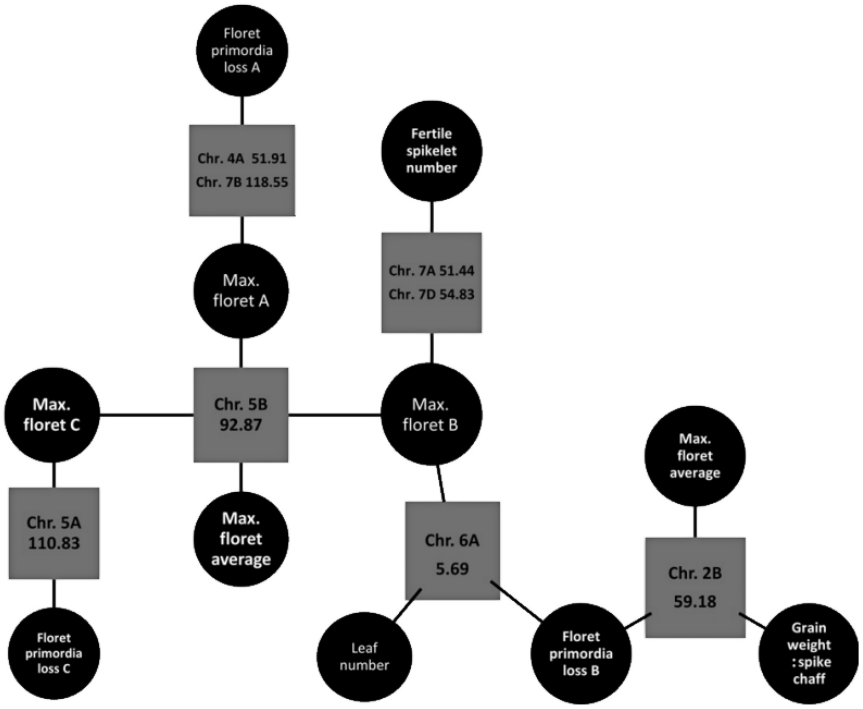


Рисунок 7. Генная сеть, основанная на обнаруженных общих QTL. QTL на хромосоме 5B является общим для признака max. floret - максимальное количество цветковых зачатков на один колосок (потенциальное количество зерен) на верхушечных (A), центральных (C) и базальных колосках (B). QTL на хромосоме 5A является общим для признаков max. floret C - максимальное количество цветковых зачатков на колосок в центральном колоске и floret primordia loss C - потеря цветковых зачатков (разница между максимальным количеством цветковых зачатков и числом зерен на колоске) на центральном колоске. QTL на хромосоме

4A и QTL на хромосоме 7B являются общими для признаков – max. floret A - максимальное количество цветковых зачатков на верхушечном колоске и floret primordia loss A - потеря цветковых зачатков на верхушечном колоске. QTL на хромосомах 7A и 7D являются общими для признаков max. floret B - максимальное количество цветковых зачатков на один колосок у основания и fertile spikelet number - число плодовитых колосков. QTL на хромосоме 6A является общим для трех признаков: max. floret B - максимальное количество цветковых зачатков у основания колоса, leaf number – число листьев на главном побеге и floret primordia loss B - потеря зачатков на колоске у основания. QTL на хромосоме 2B является общим для трех признаков: max. floret average - среднее максимальное количество цветковых зачатков на вершинных, центральных и базальных колосках; floret primordia loss B - потеря зачатков на базальном колоске; grain weight: spike chaff - соотношение массы зерна и остальной части колоса.

Результаты этой работы также показали, что даже умеренная засуха во время фазы цветения и наполнения зерна может существенно снизить урожайность пшеницы (от 58% до 92% относительно нормальных условий).

В настоящее время методы сравнительной геномики позволяют выделять, картировать и характеризовать ранее неизученные гены пшеницы, выявлять консервативные гены среди разных злаков на основе генной синтении (Gale and Devos, 1998).

Так в работе (Ma et al., 2016) изучили гомеологи гена *TaGS5*, локализованные в хромосомах 3A, 3B и 3D, тесно связанные с такими важными признаками как размер и масса зерна. В работе были идентифицированы аллели *TaGS5-3A*: *TaGS5-3A-T* и *TaGS5-3A-G*. При этом определялись такие признаки как масса 1000 зерен, длина, ширина и толщина зерна. В работе была показана значительная корреляция аллеля *TaGS5-3A-T* с увеличенным размером зерна и массой ядра. Было выявлено, что гомеологи *TaGS5* преимущественно экспрессировались в молодых колосьях и развивающихся зернах. Помимо этого, был установлен генетический маркер для различения этих двух аллелей.

В работе (Muqaddasi et al., 2019) было проведено исследование генетического контроля общего числа колосков в колосе наряду с другими признаками, такими как

длина колоса и время всходов, как составляющих урожайности для 518 элитных сортов озимой европейской пшеницы.

Полногеномный анализ ассоциаций, основанный на 39 908 маркерах однонуклеотидного полиморфизма (Single nucleotide polymorphism, SNP), выявил значимые QTL для общего числа колосков в колосе на хромосомах 2D, 7A и 7B, для длины колоса на 5A и времени всходов на 2D. В области 7A-QTL, ассоциированной с общим числом колосков, выявлено присутствие последовательности *TaAPO-A1*. Данная последовательность является ортологом гена риса *APO1*, который контролирует количество колосков на метелках. Межвидовой анализ ортологов гена *TaAPO-A1* показал, что он является высоко-консервативным геном и важен для развития цветка. Кроме того, отмечено его наличие в широком спектре наземных растений. Исследования *TaAPO-A1* по генотипам пшеницы выявили два гаплотипа в консервативном домене F-box. Маркер KASP, разработанный для идентификации полиморфного сайта, показал очень значимую связь *TaAPO-A1* с общим числом колосков в колосе, объясняя 23,2% общей генотипической дисперсии. Кроме того, аллели *TaAPO-A1* показали слабые, но существенные различия для длины колоса и урожайности зерна (Muqaddasi et al., 2019).

В вышеупомянутых работах оценка большинства фенотипических признаков растений пшеницы производилась вручную. При этом, для получения значимых результатов приходилось анализировать сотни растений. Это является трудоемким процессом, учитывая число различных оцениваемых характеристик, вычисляемых показателей, размеры оцениваемых объектов (зерен), а также число растений, задействованных в экспериментах. Для повышения производительности и качества экспериментов необходима автоматизация процессов фенотипирования, документирования, передачи и хранения полученных данных. Решением данной проблемы может служить разработка методов высокопроизводительного фенотипирования на основе анализа изображений и разработка информационных систем поддержки селекционно-генетических исследований для накопления и анализа данных.

1.2 Методы обработки биологических цифровых изображений

1.2.1 Основные процедуры подготовки и анализа изображений в биологии.

Цифровые изображения, в основном, делятся на два вида: векторные и растровые. Векторные изображения представляет собой набор математически описанных элементарных геометрических объектов (примитивов). Они легко масштабируются и эффективно используют память. Такие изображения используются как правило в компьютерной графике, в дизайне и проектировании.

Объекты реального мира могут быть представлены в векторном виде, например, при моделировании или в результате их распознавания и векторизации данных с различных датчиков, в том числе и с растровых изображений.

Растровое изображение представляет собой матрицу элементарных ячеек, хранящих в себе информацию о цвете (для цветных изображений) или интенсивности свето-потока (для черно-белых) во всех точках изображения (Форсайт и Понс, 2004). Элементы матрицы называется пикселями изображения (pixel производное от “picture element”).

Форматы представления растрового изображения в памяти компьютера могут быть различными. Они зависят от того в какой модели представлен цвет, наличия канала прозрачности (альфа-канала) и диапазона принимаемых значений для каждого из каналов - глубины цвета.

Среди обычных растровых изображений выделяют цветные, монохромные (градации серого, черно-белые) и бинарные. Монохромные изображения имеют всего один канал, хранящий значения интенсивности (градации) яркости пикселей в заданном диапазоне значений. Типичный формат представления монохромных изображений – это целочисленный массив размерности NxM, содержащий значения яркости пикселей в диапазоне от 0 до 255 (один байт). Бинарное изображение являются частным случаем монохромных. Значение интенсивности свето-потока может принимать либо 0 или 1.

Цветное изображение состоит из нескольких целочисленных массивов. Один из популярных форматов представления цветных изображений является 24-битное RGB изображение (RGB это аббревиатура английских слов Red, Green, Blue). В этом

случае изображение представлено матрицей размерности NxMx3 элементов. Каждый элемент хранит значение интенсивности одного из цветов: красного, зеленого и синего. Диапазон значений интенсивности связан с количеством выделенных бит на канал (цвет). Для 24-битного изображение интенсивность может принимать значение от 0 до 255.

Указанная выше модель описания цвета пикселей в виде интенсивности трех компонент RGB отражает технические особенности устройства записи и воспроизведения цвета в мониторах, матрицах цифровых камер и множества других устройств, но не отражает человеческое восприятие цвета. Для человека более понятны термины оттенков, яркости и насыщенности при описании цветов. Поэтому модель представления цвета через тон, насыщенность и яркость (Hue Saturation Value, HSV) является из одной наиболее удобных с точки зрения описания цветовых характеристик. Как можно судить из названия, цвет в модели задается тремя значениями:

Hue — цветовой тон. Варьируется в пределах 0—360°.

Saturation — насыщенность. Варьируется в пределах 0—100. Чем больше этот параметр, тем насыщеннее цвет, и наоборот, чем ближе этот параметр к нулю, тем ближе цвет к нейтральному серому. Иногда данный параметр называют чистотой цвета.

Value (значение) или Brightness — яркость. Задаётся в пределах 0—100.

Как и в RGB, HSV или другой цветовой модели, диапазоны значений каналов условны. По техническим особенностям, или по иным причинам, любой из них может приводится к диапазону 0—100, 0—1 или, к примеру, 1—180, как канал Hue модели HSV в библиотеке OpenCV (Kaehler and Bradski, 2016).

Приведенные описания форматов представления изображения является обзорными. На практике же существует множество других форматов представления изображения со своими особенностями.

Растровые изображения объектов реального мира как правило получают с помощью цифровых камер или сканеров. При этом в отличии от сканеров, качество

и характеристики полученных изображений в случае камер существенно зависит от множества внешних факторов, таких как число, расположение и интенсивность источников света, световой поток, температура света, наличие мерцания, расстояние объектива до объекта, а также от и параметров съемки: выдержка, диафрагма, фокусное расстояние, параметры вспышки и др.

Задача анализа изображений посвящена извлечению цифровых характеристик изображений, построению гистограмм, выделению границ биологических объектов, сегментации и т.д. Эффективность указанных методов особенно важна в случаях, когда требуется высокая точность результатов анализа изображения. Основные задачи, решаемые на стадии предварительной обработки изображений, это улучшение качества изображений, аффинные преобразования, цветовые преобразования и т.д. В настоящий момент большинство библиотек по обработке изображений реализует эти базовые функции.

При распознавании образов на изображении можно выделить следующие основные этапы: фильтрация шума, выделение областей, утоньшение или оконтуривание, векторизация, аппроксимация (Ильясова и др., 2012).

Практически любое изображение содержит искажения, например, шум или различные артефакты изображений, обусловленные условиями съемки (недостаточная освещенность/засвет) или аппаратными особенностями камеры. Методы фильтрации шума на изображении разнятся в зависимости от типа искажения. Наиболее часто встречающиеся искажения на изображении:

Методы фильтрации шума на изображении разнятся в зависимости от типа искажения. Наиболее часто встречающиеся приемом улучшения качества изображения является его размытие. Например, размытие по Гауссу. Гаусово размытие с радиусом r считается по формуле:

$$y(m,n) = \frac{1}{2\pi r^2} \sum_{u,v} e \frac{-(u^2+v^2)}{2r^2} x(m+u, n+v) \qquad (2)$$

где *x(m, n)* – значение яркости в координате *(m, n)* растра.

Такое преобразование типа растр-растр позволяет уменьшить “шум” изображения и сгладить некоторые “артефакты”, блики.

1.2.2 Сегментация, морфологические преобразования и поиск контуров

Связная область – это множество пикселей, замкнуто связанных между собой отношением «соседства». Таким образом, связная область состоит из множества точек на изображении, в котором любые две точки можно соединить друг с другом через последовательность соседей (рисунок 8).

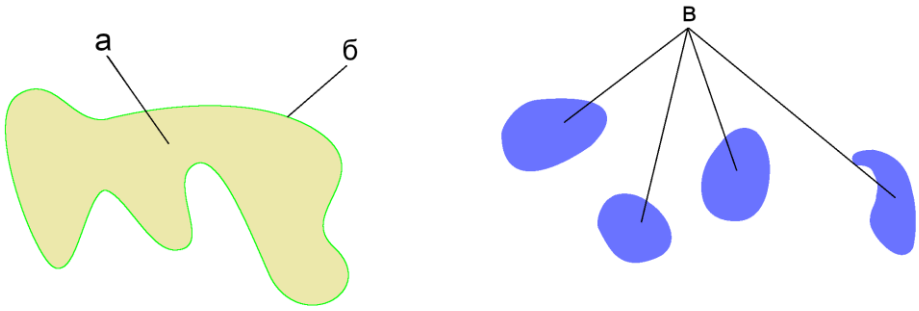


Рисунок 8. Примеры областей связности. Области состоят из пикселей одной группы (одинаковых значений цветовых каналов). а) слева, желтым – связная область б) зеленым - контур связной области в) справа, синим - несвязные области.

В зависимости от определения отношения соседства, т.е. какой из пикселей считается соседним к данному, различают 4-связность и 8-связность (рисунок 9).

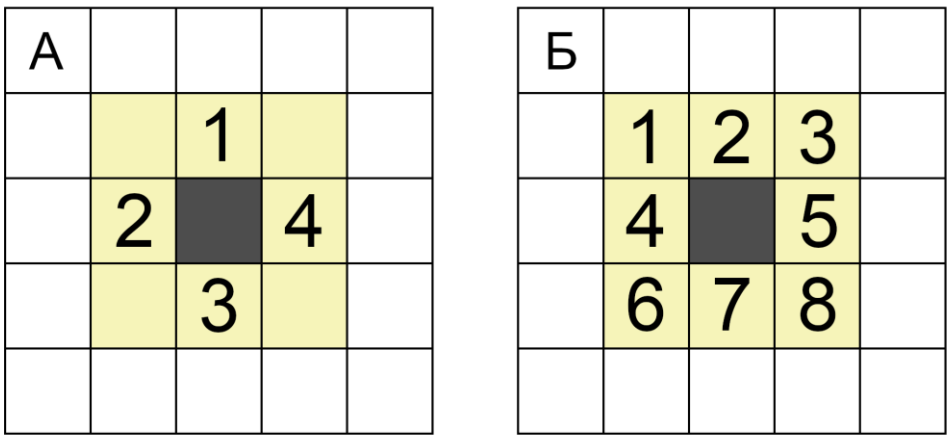


Рисунок 9. Типы связности. На рисунке изображена окрестность некоторого текущего пикселя для которого определяются пиксели-соседи. Серым цветом указан пиксель, относительно которого определяется связность (текущий пиксель). Желтым цветом указаны потенциальные пиксели-соседи. Цифрами обозначены номера соседних пикселей, которые считаются связанными с текущим пикселем при заданном типе связности. А) 4-связность. Тип связности, при котором соседними считаются пиксели, отличающиеся от текущего на единицу по одной из X или Y - координат. Т.е. пиксели непосредственно слева, справа, снизу и сверху от текущего (всего 4 пикселя). Б) 8-связность. Тип связности, при котором соседними считаются пиксели, отличающиеся от текущего на единицу по одной или обоим из X или Y - координат. Т.е. пиксели непосредственно слева, справа, снизу и сверху от текущего, а также по диагонали (всего 8 пикселей).

Контур это замкнутая или незамкнутая линия (кривая), соединяющая все непрерывные точки вдоль границы некоторой связной области. Контуры являются полезным инструментом для анализа формы, а также для обнаружения и распознавания объектов.

Сегментация - это разделение цифрового изображения на связанные подобласти. Сегментация используется для выделения объектов или интересующих областей на изображениях для дальнейшего анализа. Наиболее простой пример сегментации – пороговая бинаризация монохромного изображения, при которой все пиксели изображения на основе заранее выбранного порога интенсивности разделяются на два класса: пиксели фона и пиксели “интересующих объектов”. Например, если значение пикселя превышает пороговое – он обозначается как пиксель объекта, иначе – как пиксель фона, или наоборот, в зависимости от задачи. Порог при этом зачастую выбирается на основе параметров гистограмм. Элементы изображения соотносятся диапазону на основе минимумов, максимумов и/или экстремумов гистограмм (Shapiro and Stockman, 2001).

Ниже приведены широко применяемые методы сегментации:

- Кластеризации: формирование областей на основе выбранных центров кластеров (метод к-средних).
- Сегментация на основе выделения границ: детектор границ Canny (Canny J. A., 1986).

- Наращивание границ: сегментирование на основе сравнения средних значений яркости пикселей в соседних областях.

- Графовый метод: сегментация на основе представления изображения в виде графа: нормализованные разрезы графов (Shi and Malik, 2000), случайное блуждание (Grady, 2006), минимальный разрез (Wu and Leahy, 1993), изопериметрическое разделение (Grady and Schwartz, 2006) и сегментация с помощью минимального остовного дерева (Zahn, 1971).

- Метод водораздела: изображение рассматривается как топографическая поверхность. Пиксели, имеющие наибольшую абсолютную величину градиента яркости, соответствуют линиям водораздела, которые представляют границы областей. Локальные минимумы яркости образуют основы сегментов. Пиксели распределяются по сегментам на основе того, к какому локальному минимуму направлен его градиент яркости (Roerdink and Meijster, 2000).

1.2.2.1 Морфологические преобразования

Для объяснений морфологических операций будут использоваться бинарные изображения, т.е. изображения, пиксели которых условно принимают значения 0 либо 1. Условность заключается в том, в зависимости от реализации, в качестве единицы может быть 255 или любое другое значение, отличное от нуля. Главное, что число возможных состояний пикселя в таких изображениях равно двум, и под нулем принимается фон (задний план), а под “не нулем” принимается объект переднего плана.

Морфологические преобразования - это простые операции, обычно выполняющиеся на двоичных изображениях. Они производятся с помощью так называемого ядра преобразования – квадратной матрицы, которая определяет характер операции. Типичные морфологические операторы - это эрозия, расширение и их разновидности, такие как открытие, закрытие, морфологический градиент и др.

Основная идея эрозии похожа на эрозию почвы, она размывает границы объекта переднего плана (рисунок 10). Ядро проходя по изображению, как в свертке, вычисляет результирующий пиксель как 1, если все пиксели под ядром равны 1, в

противном случае он равен 0. Таким образом, все пиксели вблизи границы будут отбрасываться в зависимости от размера ядра. При этом размер белой области (объекта переднего плана) на изображении уменьшается. Операция полезна для удаления небольших белых шумов.

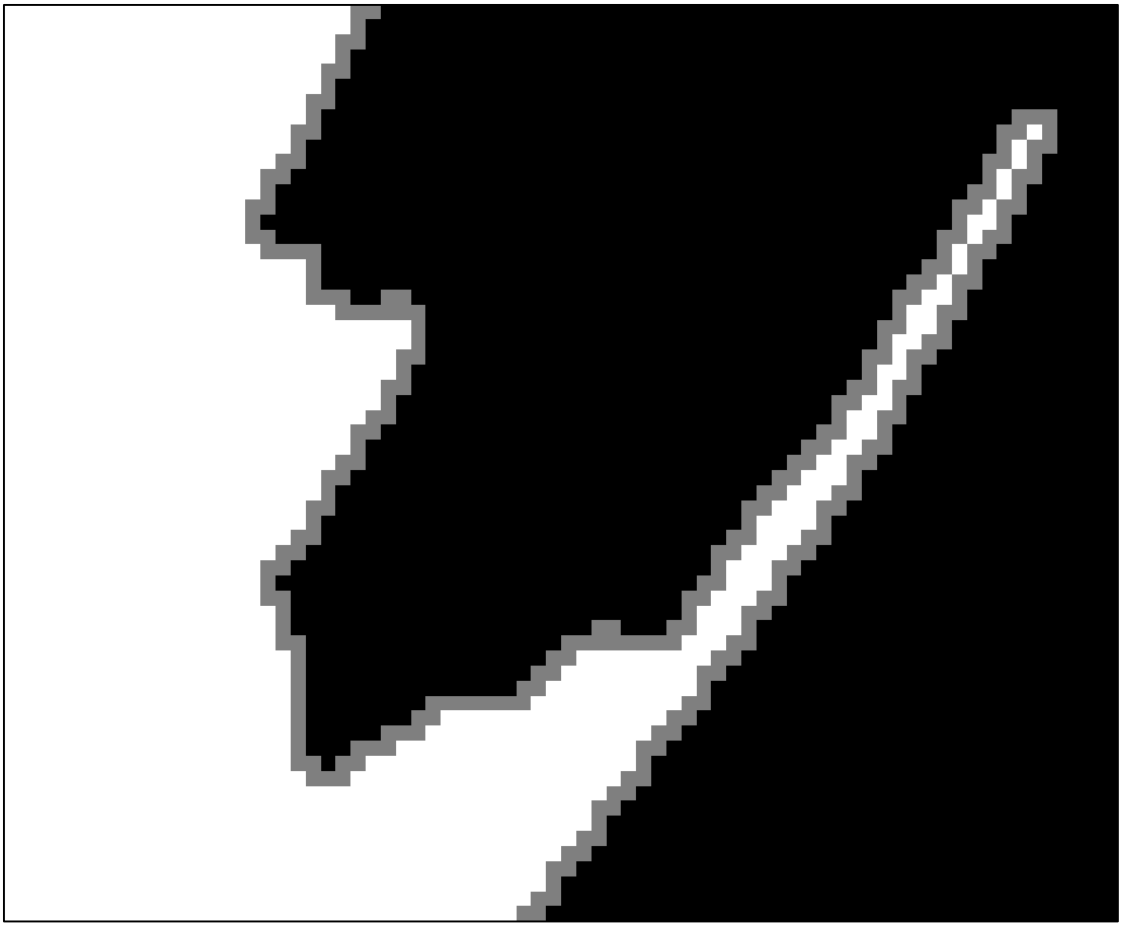


Рисунок 10. Пример морфологического преобразования эрозии. Белым цветом отмечены пиксели объекта. Серым цветом отмечены граничные с фоном пиксели объекта, которые будут удалены (заменены цветом фона) преобразованием. Черным цветом – фон.

Расширение - это противоположность эрозии. Результирующий пиксель в данном преобразовании будет равен “1”, если хотя бы один пиксель под ядром равен “1”. Таким образом, белая область (объект переднего плана) на изображении увеличивается.

Обычно в таких случаях, как удаление шума, за эрозией следует расширение, потому что эрозия удаляет белые шумы, но также удаляют точки объект переднего плана, а расширение восстанавливает исходные границы объекта. Такое преобразование называют открытием.

Закрытие — это расширение с последующей эрозией. Метод полезен при закрытии небольших отверстий внутри объектов переднего плана или маленьких черных точек на объекте.

Ещё одно полезное морфологическое преобразование – “скелетизация” или “утоньшение” (от англ. “thinning”). Данное преобразование применяется для выделения остовов (скелетов) областей (рисунок 11). Основная идея данных алгоритмов состоит в интерактивном удалении внешних слоев и контурных точек объектов до тех пор, пока на изображении не останутся только точки остова (скелета) (Ильясова и др., 2012).

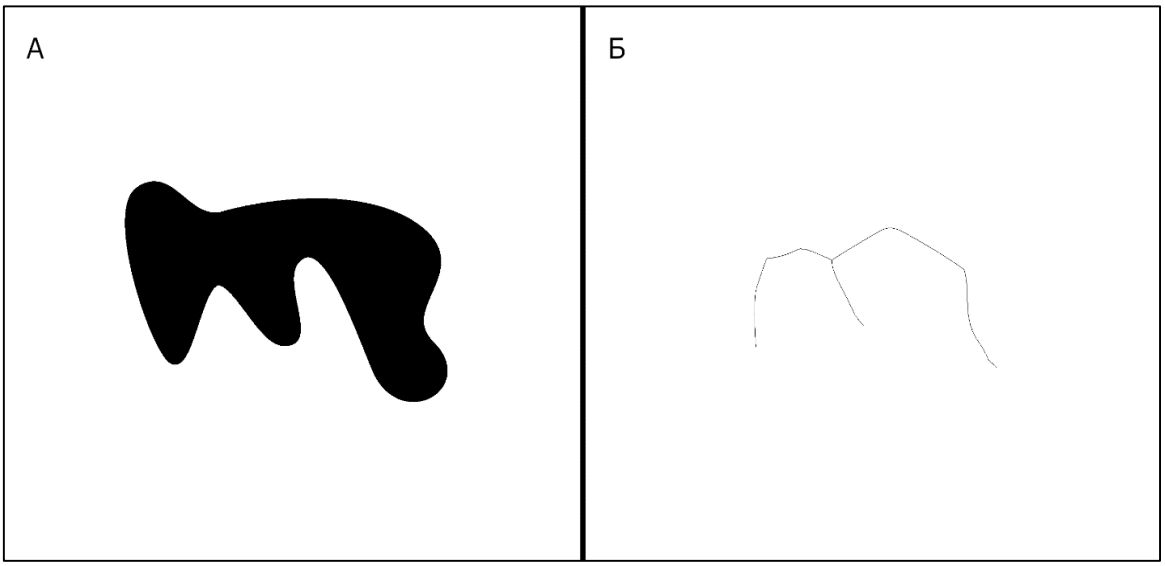


Рисунок 11. Скелетизация. А) Связная область (отпечаток). Б) Результат скелетизации. Белым цветом отмечен фон на изображении.

1.2.3 Задачи оптимизации

В области обработки изображений нередко возникают различного рода оптимизационные задачи, когда требуется найти наиболее оптимальные параметры функции или алгоритма. Оптимизация — это задача нахождения экстремума (минимума или максимума) целевой функции в некоторой области конечномерного векторного пространства, ограниченной набором линейных и/или нелинейных равенств и/или неравенств. Теорию и методы решения задачи оптимизации изучает математическое программирование.

Существующие методы поиска включают детерминированные алгоритмы; эвристические/стохастические; задачи дискретного программирования; задачи целочисленного программирования; задачи линейного/нелинейного программирования.

Примером эвристического алгоритма поиска служит генетических алгоритм (Whitley, 1994). Генетический алгоритм — это эвристический алгоритм поиска, используемый для решения задач оптимизации и моделирования путём последовательного подбора, комбинирования и вариации искомых параметров с использованием механизмов, напоминающих биологическую эволюцию.

Отличительной особенностью генетического алгоритма является акцент на использование оператора «скрещивания», который производит операцию рекомбинации (кроссинговер) особей-решений, роль которой аналогична роли скрещивания в живой природе. Особями-решениями могут быть наборы параметров.

При решении оптимизационных задач генетическим алгоритмом, моделируется популяция особей-решений, вы из них выбираются лучшие кандидаты и выполняется скрещивание. После чего на основе полученных новых особей решений генерируется новая популяция.

1.2.4 Задача распознавания образов

Образ (класс) — это совокупность данных о реальном или абстрактном объекте, позволяющая выделять его из всего множества анализируемых данных. Распознавание образов – это идентификация и классификация объектов, которые характеризуются конечным набором некоторых свойств и признаков. Задача распознавания образов является одной из основных в большинстве интеллектуальных систем, таких как машинное зрение (Zayas et al., 1989). Цель таких систем состоит в получении изображения и составлении его описания. В целом задачу распознавания можно разделить на два этапа. Первый это детектирование (обнаружение) объектов на изображении. Второй этап – это классификация.

Задача детектирования образа состоит в определении всех областей изображения, возможно содержащих интересующий объект.

Задача классификации состоит в определении, к какому конкретному классу принадлежит обнаруженный объект. Роль исходной информации для классификации образа играет понятие вектор признаков. Признак – это некоторое качественное или количественное измерение распознаваемого объекта произвольной природы. Вектор признаков – это множество признаков одного образа. В рамках задачи распознавания считается, что каждому образу ставится в соответствие единственное значение вектора признаков и наоборот: каждому значению вектора признаков соответствует единственный образ.

Классификация в системах распознавания основывается на образах, правильная классификация которых известна. Можно считать, что все объекты или явления разбиты на конечное число классов. Задача распознавания, на конечном этапе, состоит в том, чтобы отнести новый распознаваемый объект к какому-либо классу (Местецкий, 2008).

Классификатором или решающим правилом классификации называется правило отнесения образа к одному из классов на основании его вектора признаков. Практическая разработка системы классификации осуществляется по схеме решения следующих задач:

- Задача генерации признаков – выбор признаков, описывающих образ наиболее полно.
- Задача селекции признаков – отбор наиболее эффективных в классификации признаков.
- Задача построения классификатора – выработка правила классификации, по которому на основании вектора признаков будет осуществляться классификация.
- Задача оценки системы – выбор правила для количественной оценки правильности системы классификации.

Задача классификации нередко решается с помощью нейронных сетей, так как число признаков и параметров влияния их на конечный результат зачастую достигает нескольких десятков или сотен. При таких условиях решение задачи

построения классификатора становится слишком сложной. Тем не менее, множество различных работ по распознаванию образов нередко показывают результаты, указывающие на эффективную применимость более простых классификаторов с адекватно подобранным вектором признаков. Подход с простыми классификаторами является приемлемым для ряда задач распознавания, особенно в системах, где производительность является критическим параметром.

1.2.5 Применение программных средств для анализа данных фенотипа растений

Для определения взаимосвязи фенотипа и генотипа пшеницы применяются различные методы, такие как:

- морфометрический анализ (сравнение объектов по их форме/размерам/площади);
- статистический анализ (определение природы случайной совокупности величин, выявления взаимосвязей и структур в данных);
- компьютерно-экспериментальные подходы (применение компьютерных подходов для симуляции биологических экспериментов *in silico*), основанные на методе картирования генов, контролирующих QTL.

В работе (Gegas et al., 2010), посвященной анализу морфологии зерен пшеницы идентифицировались основополагающие гены, связанные с размером зерен и их формой путем поиска QTL. Для измерений ширины, длины, площади зерна и других параметров использовался анализатор MARVIN. После были рассчитаны отношение длины к ширине зерна и различные характеристики, описывающие плотность зерна и его форму. Статистический анализ количественных данных проводили с использованием программы Genstat V 11. Так же был проведен дисперсионный анализ для оценки относительного генетического вклада в изменение характеристик для каждой популяции и для каждого года. Для картирования и идентификации QTL использовалось программное обеспечение MapQTL 5.0. Широкомасштабный численный анализ позволил определить генетическую основу фенотипического разнообразия в морфологии зерен пшеницы, зафиксировать вариации размеров и форм зерен в многочисленных популяциях. Результаты такого анализа выявили, что

размер и форма зерна в большинстве своем являются независимыми признаками у ранних и современных сортов.

Применение программно-аппаратных средств развивается также и для других растений. Примерами этому служат:

- Комплекс PlaRoM (plant root monitoring platform), который позволяет осуществлять мониторинг роста корней растений (Yazdanbakhsh and Fisahn, 2009).
- Метод автоматического определения количественных характеристик опушения листьев растений на основе анализа цифровых микроизображений их сгибов (Genaev et al., 2012).
- Метод позволяющий производить оценку биомассы различных злаков (Golzarian et al., 2011).
- Метод быстрого определения размеров семян у арабидопсиса (Herridge et al., 2011).
- Метод объективной оценки цвета плодов фруктов и овощей даёт точную количественную оценку различных характеристик цвета (яркость, насыщенность, цветовой тон) и монотонности цвета (Darrigues et al., 2008).

Таким образом, различные методы морфометрии и статистического анализа, а также компьютерно-экспериментальные подходы успешно применяются для определения статистической связи между хозяйственными признаками растений (продуктивность, устойчивость к стрессам) и локусами в геноме (генами или локусами количественных признаков, QTL).

1.3 Библиотеки анализа цифровых изображений

В области биоинформатики для решения задач обработки и анализа изображений широко используются различные универсальные программные пакеты. Рассмотрим наиболее распространенные из них.

1.3.1 Matlab

Matlab - пакет прикладных программ для решения задач технических вычислений и одноимённый язык программирования, используемый в этом пакете (Дьяконов, 2002). Matlab включает пакет прикладных программ Image Processing Toolbox, который включает ряд встроенных функций, реализующих наиболее распространенные методы обработки изображений, такие как: геометрические преобразования (кадрирование, изменение размеров, поворот), вычисление статистических характеристик изображения (гистограммы, средние значения, корреляции), функции фильтрации и корректировки изображений (фильтры Лапласа и Гаусса, эквализация, коррекция динамического диапазона), морфологические преобразования и сегментация.

Пакет широко распространен среди инженерных и научных работников, так как предоставляет пользователю большое количество функций для анализа данных, покрывающие практически все области математики, в частности:

- Матрицы и линейная алгебра — алгебра матриц, линейные уравнения, собственные значения и векторы, сингулярности, факторизация матриц и другие.
- Многочлены и интерполяция — корни многочленов, операции над многочленами и их дифференцирование, интерполяция и экстраполяция кривых и другие.
- Математическая статистика и анализ данных — статистические функции, статистическая регрессия, цифровая фильтрация, быстрое преобразование Фурье и другие.
- Обработка данных — набор специальных функций, включая построение графиков, оптимизацию, поиск нулей, численное интегрирование (в квадратурах) и другие.
- Дифференциальные уравнения — решение дифференциальных и дифференциально-алгебраических уравнений, дифференциальных

уравнений с запаздыванием, уравнений с ограничениями, уравнений в частных производных и другие.

- Разреженные матрицы — специальный класс данных пакета MATLAB, использующийся в специализированных приложениях.
- Целочисленная арифметика — выполнение операций целочисленной арифметики в среде MATLAB.

Matlab имеет средства для поддержки разработки алгоритмов на собственном языке (в частности поддержка объектно-ориентированного программирования), функции визуализации данных (в том числе построение 3-х мерных графиков), компилятор, поддержку web-сервисов (SOAP и WSDL).

1.3.2 Среда R для анализа изображений

R – это интерпретируемый язык программирования и свободная программная среда, под одноименным названием (Team et al., 2013). Несмотря на то, что R в основном используется для статистической обработки данных, математических вычислений и работой с графикой, он также применяется для анализа изображений посредством специальных пакетов. Так, например, пакет EBImage для R предоставляет функции для обработки и анализ изображений в области клеточной биологии на основе данных микроскопии (Pau et al., 2010). EBImage предлагает инструменты для сегментации клеток и извлечения количественных клеточных дескрипторов, а также набор универсальных функций для чтения, записи, обработки и анализа изображений. Пакет R LeafArea является удобным и автоматизированным инструментом для быстрого цифрового анализа отсканированных изображений листьев (Katabuchi, 2015). Пакет позволяет измерять площадь и оценивать ряд других функциональных характеристик листа. Пакет Phenopix, для извлечения фенологической информации из покадровой цифровой фотографии растительного покрова (Filippa et al., 2016). Пакет позволяет рисовать области интереса (ROI) на изображении; изображать траектории показателей зелени и вычислять кривую сезонных траекторий; извлекать фенофазы (фенологические пороги) и т.д.

1.3.3 Библиотека OpenCV

Современные библиотеки обработки и анализа изображений с открытым исходным кодом позволяют производить высокопроизводительные вычисления для задач компьютерного зрения. Пример такой библиотеки – OpenCV, реализованной для языков программирования C++, Java, Python и др. (Dawson-Howe, 2014). Библиотека портирована на все современные операционные системы, включая мобильные, такие как Android и iOS. OpenCV может свободно использоваться в академических и коммерческих целях и распространяется в условиях лицензии BSD.

Библиотека реализует все основные функции по обработки и анализу изображений. Помимо реализации базовых структур и вычислений (математических функций, генераторов случайных чисел, методов линейной алгебры, DFT, DCT) и основных функций обработки изображений (фильтрации и корректировки, геометрических и морфологических преобразований, преобразование цветовых пространств, сегментации и т.д.), библиотека также включает методы распознавания образов на изображении, анализ движения и отслеживание объектов, модели машинного обучения и многое другое.

1.3.4 ImageJ

ImageJ — программное обеспечение на языке Java с открытым исходным кодом для анализа и обработки изображений, распространяющееся без лицензионных ограничений. ImageJ широко применяется в биомедицинских исследованиях, астрономии, географии и других дисциплинах, связанных с анализом изображений, в качестве альтернативы проприетарному ПО. Плагины сторонних разработчиков охватывают широкий круг задач анализа и обработки изображений. Архитектура плагинов ImageJ и встроенная в программу система разработки делает эту платформу весьма популярной для работы и преподавания анализа и обработки изображений. Программа распространяется также в виде дистрибутива, ориентированного на работу с медико-биологическими изображениями - Fiji. Это собранный пакет обработки изображений, включающий множество плагинов, которые облегчают научный анализ изображений.

1.3.5 Scikit-image

Scikit-image - библиотека обработки изображений с открытым исходным кодом для языка программирования Python. Она включает в себя алгоритмы и утилиты для использования в исследовательских, образовательных и промышленных приложениях. Scikit-image реализует такие алгоритмы обработки изображений, как сегментация, геометрические преобразования, манипулирования цветовым пространством, фильтрация, морфологические преобразования, распознавание образов и т.д. Библиотека разработана для взаимодействия с числовыми и научными библиотеками Python NumPy и SciPy.

1.4 Методы фенотипирования колосьев и зерен пшеницы

1.4.1 Фенотипирование колосьев

Наиболее простым способом оценки характеристик колоса является визуальная оценка (сопоставление с шаблоном по типу), измерение размеров с помощью ручных инструментов, а также ручной подсчет зерен в колосе и их взвешивание. Такой способ позволяет классифицировать колос и подробно описать его особенности. Указанные характеристики определяются вручную, что является достаточно трудоемким процессом, принимая во внимание необходимость обработки большого количества данных в современных селекционно-генетических экспериментах. Следует также учесть, что результаты таких измерений, как правило, документируются вручную, без использования современных технологий хранения данных.

Такие характеристики могут быть измерены и задокументированы с применением методов высокопроизводительного массового фенотипирования растений и современных технологий хранения и передачи данных.

В особую категорию можно отнести так называемые неразрушающие методы изменения - измерения, при проведении которых не требуется срывать/срезать и как-либо повреждать растение. Так, в работе (Kun et al., 2010) был предложен метод неразрушающего измерения характеристик колоса. Измерялись такие характеристики, как количество остей, средняя длина остей и длина колоса, на основе обработки и анализа изображений. Полученные параметры легли в основу

трехслойной нейронной сети для классификации пшеницы на 4 сорта. Точность классификации составила 88% на 240 изображениях.

1.4.2 Анализ формы зерен

Ранее были предложены различные методы для эффективной морфометрии семян растений с использованием технологий обработки изображений (Granitto et al., 2005; Pourreza et al., 2012; Tanabata et al., 2012). Большинство из них были реализованы с использованием программного обеспечения настольного персонального компьютера (ПК) для анализа изображений зерен на светлом фоне, полученном с использованием цифровой камеры или сканера (Herridge et al., 2011; Tanabata et al., 2012; Whan et al., 2014). Данные подходы позволяют пользователям: оценивать большое количество морфометрических параметров зерна, описывающих форму и цвет (Bai et al., 2013); облегчают решение задачи идентификации сортов злаковых путем анализа изображений зерен (Wiesnerová and Wiesner, 2008; Chen et al., 2010; Zapotoczny, 2011); помогают определять содержание влаги в семенах и прогнозировать выход манной крупы для твердой пшеницы (Novaro et al., 2001; Tahir et al., 2007). В работе Дуана и его соавторов (Duan et al., 2011) была предложена автоматическая система для высокопроизводительного анализа характерных черт риса, влияющих на урожайность: общее количество колосков, количество непустых колосков, вес 1000 зерен, длина и ширина зерна и др. Роуссел и др. (Roussel et al., 2016) предложил подробный анализ формы и размера семян. Они использовали 3D-реконструкцию поверхности из силуэтов нескольких изображений, полученных путем вращения зерна перед цифровой камерой.

Позже, этот метод был реализован на роботизированной платформе phenoSeeder (Jahnke et al., 2016), которая была разработана для высококачественного измерения основных биометрических признаков и массы зерен, из которых также рассчитывается и их плотность. Стрэндж и др. (Strange et al., 2015) использовали рентгеновскую компьютерную томографию для определения формы зерна у растений. Инженерные оборудование для морфометрии зерна демонстрируют высокую производительность и точность, однако они установлены в ограниченном количестве исследовательских лабораторий. Поэтому на сегодняшний момент по-

прежнему существует потребность в недорогих высокопроизводительных методах анализа морфометрических параметров зерна (Whan et al., 2014).

Одно из существующих программных средств, обеспечивающее высокопроизводительное измерение семян является разработанное японскими исследователями компьютерная программа SmartGrain (Tanabata et al., 2012). SmartGrain использует изображения зерен, полученные при помощи сканера, и автоматически распознает на них все семена, обнаруживает их контуры, а затем вычисляет длину, ширину, площадь семян, длину периметра, и другие параметры. Такое программное обеспечение может быть использовано в анализе QTL. Продемонстрированные результаты работы со SmartGrain показывают надежность и эффективность использования данного подхода для использования в генетическом анализе.

SmartGrain использует алгоритм сегментации (Tanabata T. et al., 2010) для идентификации контуров зерен, а также, анализ морфологии для удаления ости и стебля. Приложение автоматически обнаруживает перекрытие семян на изображении и исключает такие объекты из анализа, а также автоматически удаляет ости и плодоножки. Для обнаружения зерен, определения степени их перекрытия и порога удаления ости и цветоножки программа использует набор управляющих параметров. Авторы сообщают, что анализ изображения с разрешением 600 точек на дюйм занимает несколько минут. Результаты могут быть экспортированы в CSV файл, который поддерживается различными программами для работы с электронными таблицами (например, Microsoft Excel).

1.4.3 Морфометрия растений при помощи мобильных устройств

Крупномасштабные селекционные эксперименты требуют обработки существенных объемов фенотипических данных, часто в полевых условиях и, таким образом, без доступа к настольным компьютерам и сканерам. В таком случае использование цифровой фотокамеры - приемлемый вариант, но изображения впоследствии должны быть скопированы на ноутбук или ПК, соответствующе подготовлены и затем проанализированы специальным программным

обеспечением. Зачастую существенную часть времени при этом занимает сортировка, сопоставление и переименование полученных файлов.

В современные мобильные устройства (смартфоны и интернет-планшеты) встроены цифровые камеры с высоким разрешением. Мобильные устройства имеют многоядерные процессоры с достаточной вычислительной мощностью для обработки и анализа изображений. Эти функции позволяют пользователям захватывать и обрабатывать изображения везде, где это необходимо. Для морфометрии органов растений разработан ряд приложений для мобильных устройств. Например, приложение Leafsnap (Kumar et al., 2012) способно идентифицировать виды растений в реальном времени на основе изображений их листьев. Для этого пользователь фотографирует растительный лист с помощью мобильного устройства и отправляет изображения с камеры на удаленный сервер, где они и обрабатываются. LeafDoctor (Pethybridge and Nelson, 2015) – это еще одно мобильное приложение, которое позволяет полуавтоматическим способом оценивать процент поражения тканей растения различными заболеваниями на основе изображений листьев. Мобильные устройства также могут служить эффективными инструментами для оценки цвета почвы (Gómez-Robledo et al., 2013).

1.5 Базы данных и онтологии в области феномики, селекции и генетики растений

Современные эксперименты по генетике и селекции включают анализ тысяч растений, их фенотипов и генотипов. Накопление огромных масштабов информации в биологии сопровождалось повсеместным распространением молекулярно-биологических баз данных. Эти базы были ориентированы на аккумулирование данных определенного вида: геномных и транскриптомных последовательностей, экспрессирущихся последовательностей тегов (Expressed Sequence Tag, EST), белков, структур генов и описания их функций, информации о метаболических путях и метаболитах, и т.д. Данные накапливались для различных видов организмов, в том числе для растений.

Большой опыт в биоинформатике по хранению и обработке данных был получен в процессе работы с молекулярно-генетическими данными (Marx, 2013). Это, прежде

всего, данные, которые получены в результате работ по секвенированию генома (геномика) (O'Driscoll et al., 2013), анализу экспрессии генов (RNA sequencing, RNA-seq), анализу геномной регуляции (ChIP-sequencing, CHiP-seq), геномных вариаций (Single nucleotide variants, SNV). Рост этих данных во многом обусловлен появлением современных высокопроизводительных технологий чтения геномных последовательностей. Указанные технологии позволили относительно недорого извлекать информацию о последовательностях ДНК и их вариациях в геноме. Со времени секвенирования первого генома человека (начало пост-геномной эры) решению вопросов хранения и обработки геномных данных были посвящены огромные усилия информатиков и биологов. Это позволило создать полезные для биологов ресурсы для работы с геномными данными: Ensembl (www.ensembl.org) - база геномных данных и аннотаций (Flicek et al., 2011); Expression Atlas (https://www.ebi.ac.uk/gxa/home) - база данных экспериментов по экспрессии генов и т.д. Однако, в то время как объем данных рос огромными темпами, их аннотирование, обработка и анализ существенно отставали.

1.5.1 Феномика

Для успешного решения биологических задач информации только лишь по геномным данным оказывается недостаточно. Во многом успех зависит от того, насколько полно информация о генотипах организмов будет интегрирована с информацией о фенотипах (наблюдаемых биологических признаках, заболеваниях и т.п.). Именно поэтому одно из бурно развивающихся направлений в современной биологии на стыке с информационными технологиями – разработка эффективных технологий быстрого, точного и массового определения фенотипических признаков организмов (высокопроизводительная феномика) (Hancock, 2014). Этот раздел науки определяет протоколы сбора, стандарты описания результатов лабораторных экспериментов, описание изменений в физиологии.

Методы феномики получили бурное развитие в биологии растений, поскольку именно в этой области сбор фенотипических данных и оценка фенотипических параметров представляет большую трудность в связи с необходимостью работы в полевых условиях и оценки большого числа параметров фенотипа (Furbank and Tester, 2011). Задачи феномики в настоящее время представляют собой вызов для

информатики. Обусловлено это необходимостью интеграции фенотипических данных с геномными, и учетом информации об окружающей среде, которая определяет не менее 50% вариаций фенотипа. Помимо этого, растет потребность в разработке новых технологий фенотипирования, в том числе и на основе анализа двухмерных (Генаев и др., 2012; Eliceiri et al., 2012), трехмерных (Paproki et al., 2012) изображений, а также высокотехнологичных сенсорных технологий (Busemeyer et al., 2013). Особенно остро для феномики растений стоит проблема интерпретации многопрофильных данных в связи с необходимостью описания параметров окружающей среды и экологии (Madin et al., 2007). Сбор и интеграция между собой экологических данных является нетривиальной задачей, т.к. охватывает чрезвычайно широкий спектр типов данных, структур и семантических понятий.

Помимо этого, чрезвычайно важна потенциальная значимость интеграции данных о фенотипе из нескольких источников. К примеру, из различных лабораторий, с вариациями методов измерения подобных фенотипов, нескольких популяций, или штаммов конкретного организма.

1.5.2 Онтологии и их применение для решения задач биоинформатики

По мере накопления информации в базах данных, возрастала потребность в обработке молекулярно-биологической информации растений различных видов для выяснения функциональных аспектов биологии растений и проведения сравнительного геномного, транскриптомного и протеомного анализа.

Поэтому, исследователи все больше нуждались в междисциплинарном подходе для более ясного понимания биологических процессов генной экспрессии при формировании фенотипа. Проведение сравнительного геномного анализа требует интеграции данных из нескольких источников. Для более эффективного использования биологических баз данных и знаний, которые они собирают, различные ресурсы должны быть интегрированы таким образом, чтобы полученный результат был наиболее удобен и полезен для биологов. Однако термины, используемые для описания сравниваемых объектов внутри и между базами данных, могут сильно отличаться, что препятствует точным, успешным и унифицированным запросам информации между различными базами данных.

При анализе фенотипов длительное время общепринятой практикой была разработка каждый раз новых протоколов проведения экспериментов.

Это приводило к тому, что каждое исследование документировалось уникальным образом, что затрудняло сравнение результатов экспериментов, проводимых в различных лабораториях.

Другая проблема заключается в недостатке стандартизации названий характерных фенотипических признаков растений и методов их измерения, особенно применяемых в феномике (Hancock, 2014).

Одним из решений данных проблем является разработка и применение стандартов аннотаций, таких как структурированные управляемые словари, организованные в онтологии. Онтология, это классификационная методология для формализации знаний о предмете исследования в структурированной форме. Как правило такая формализация полезна в электронных базах данных - она необходима для точной и последовательной документации сущностей, их свойств и функций (как например, продукты и функции генов, генная структура, характерные черты растений, их фенотипы, стадии развития и т.д.) (Подколодный и Подколодная, 2016).

В биоинформатике, онтологии - это формальное представление множества концепций и связей между ними внутри определенной дисциплины или области знаний. Онтологии предоставляют разделяемый и управляемый словарь, который может быть использован для моделирования области знания в терминах типов объектов или концепций, и их свойств, и взаимосвязей.

Большая часть генов, определяющих основные биологические функции, является общей для всех эукариот. Знание биологической роли общих белков в одном организме часто может быть передано другим организмам (Botstein et al., 2000). Аналогично, фенотипические признаки у разных организмов могут иметь общую основу и выражаться схожими терминами. Онтологии позволяют формировать общие, структурированные и динамически контролируемые словари для описания фенотипов и служат важным инструментом для эффективной аннотации и всестороннего поиска данных фенотипа (Smith and Eppig, 2012).

Несмотря на общие исследуемые признаки растений, исследователи получают очень разнородные данные в рамках экспериментов, проведенные по различным протоколам и в различных условиях. Отсутствие общего и структурированного словаря, используемого независимыми сторонами для описания своих наборов данных, является преградой для объединенного сравнительного анализа. В этой связи, онтологии помогают решать проблему интеграции данных между различными поставщиками данных (Matteis et al., 2013), обеспечивая концептуальную основу областей знаний и способствуя коммуникации исследователей в разных областях.

Помимо этого, онтологии обеспечивают поддержку сложных, автоматизированных, интеллектуальных запросов данных и их сравнение, когда структура хранящихся данных не фиксирована и требуется более гибкая логика работы с запросами таких данных. Онтологии могут служить схемами для баз данных. На данный момент научными сообществами уже создаются и развиваются системы на основе словарей и онтологий, позволяющие связывать данные экспериментов с их полным и формальным описанием.

1.5.3 Gene Ontology

Наиболее известная биологическая онтология это Gene Ontology (Gene Ontology Consortium et al., 2011), которая описывает гены и генные продукты некоторых модельных организмов. Проект Gene ontology (GO) - это база данных и информационный ресурс, предоставляющий множество структурированных, управляемых словарей и классификаций для описания ключевых областей молекулярной и клеточной биологии. Проект включает описания свойства генов, генных продуктов и биологических последовательностей, которые свободно доступны для аннотирования.

База данных GO интегрирует словари, поддерживает аннотации и предоставляет доступ к этой информации в нескольких различных форматах. Web-ресурс GO также предоставляет доступ к обширной документации о проекте GO и ссылки на работы, где используются данные GO для функционального анализа. GO предоставляет

согласующиеся дескрипторы генных продуктов для различных баз данных, а также стандартизацию последовательностей и их особенностей.

Множество баз данных, в том числе и модельных организмов, используют ссылки на термины GO наряду с собственными наборами данных, аннотациями, со ссылками на объекты других баз данных. Аннотации при этом снабжаются ссылками на источник, которыми могут быть: литературные источники, иные базы данных, результаты сторонних исследований и т.д., и указывают виды доказательств, которые предоставляет источник для более полного описания взаимосвязи между генным продуктом и термином GO.

1.5.4 PlantOntology

База данных онтологии растений, основанная на онтологии генов (Ilic et al., 2007) и онтологиях модельных видов растений (арабидопсис, рис, кукуруза). PlantOntology позволяет пользователям приписывать атрибуты структуры растений и стадий развития типам данных, таких как гены и фенотипы, для обеспечения семантической основы, необходимой для преодоления понятийного барьера, разделяющего базы данных характерных черт растений и проведения сравнительных исследований. Онтология содержит множество терминов, и их связи с генами и их продуктами по арабидопсису, рису и кукурузе. С момента своего основания, консорциум выпустил три большие онтологии:

1) Онтология структуры растений - является одной из верхнеуровневых онтологий проекта, которая описывает морфологические и анатомические структуры и включает в себя органы и системы органов, тканей и типы клеток;

2) Онтология этапов роста растения, которая описывает этапы роста организма (такие как “прорастание”, “рост розетки”, “цветение” и “старение”) и охватывает вегетативный и репродуктивный жизненный цикл всего растения;

3) Онтология этапов развития структуры растения, которая предназначена для описания этапов развития отдельных общих структур растений, а именно цветка, листьев, плодов, соцветия, корней и семян (Avraham et al., 2008).

1.5.5 Crop ontology

Для того, чтобы облегчить обмен данными между различными биологическими базами, были собраны существующие публичные словари понятий от различных исследователей культурных растений (Shrestha et al., 2012). После, между терминами были определены взаимосвязи и созданы онтологии на их основе. Таким образом была создана база данных онтологии культур Crop ontology (CO). Онтологии культур обеспечивают наборы словарей для некоторых экономически важных видов растений, таких как: пшеница, соя, рис, картофель, кукуруза и другие. Контролируемые словари CO используются в управлении агрономическими базами данных нескольких центров A Global Agricultural Research Partnership (CGIAR). Использование терминов онтологии для описания агрономических фенотипов и точных отображений этих описаний в базу данных становится важным этапом в исследованиях фенотипических и генотипических особенностей видов растений.

1.6 Статистический анализ данных

Ранее упомянутый пакет прикладных программ для решения задач технических вычислений Matlab (Дьяконов, 2002), в первую очередь широко распространен среди инженерных и научных работников, так как предоставляет пользователю большое количество функций для анализа данных, покрывающие практически все области математики, в частности:

• Матрицы и линейная алгебра — алгебра матриц, линейные уравнения, собственные значения и векторы, сингулярности, факторизация матриц и другие.

• Многочлены и интерполяция — корни многочленов, операции над многочленами и их дифференцирование, интерполяция и экстраполяция кривых и другие.

• Математическая статистика и анализ данных — статистические функции, статистическая регрессия, цифровая фильтрация, быстрое преобразование Фурье и другие.

• Обработка данных — набор специальных функций, включая построение графиков, оптимизацию, поиск нулей, численное интегрирование (в квадратурах) и другие.

• Дифференциальные уравнения — решение дифференциальных и дифференциально-алгебраических уравнений, дифференциальных уравнений с запаздыванием, уравнений с ограничениями, уравнений в частных производных и другие.

• Разреженные матрицы — специальный класс данных пакета MATLAB, использующийся в специализированных приложениях.

• Целочисленная арифметика — выполнение операций целочисленной арифметики в среде MATLAB.

Matlab имеет средства для поддержки разработки алгоритмов на собственном языке (в частности поддержка объектно-ориентированного программирования), функции визуализации данных (в том числе построение 3-х мерных графиков), компилятор, поддержку web-сервисов (SOAP и WSDL).

Для статистического анализа биологических, и не только, данных широко используется программный пакет Statistica для статистического анализа, разработанный компанией StatSoft (Petrie, 2002). Пакет имеет различные сборки в зависимости от нужд пользователя и включают широкий спектр мощных аналитических инструментов, реализующих функции анализа, визуализации и управления данными с привлечением статистических методов, а также прогнозирование, нейросетевые вычисления, контроль качества, “data mining”.

Пакет обладает широкими графическими возможностями, позволяет выводить информацию в виде различных типов графиков (включая научные, деловые, трёхмерные и двухмерные графики в различных системах координат, специализированные статистические графики — гистограммы, матричные, категорированные графики и др.), все компоненты графиков настраиваются. Данный программный пакет удобен в использовании, и не требует значительного времени на обучение, тем не менее он является проприетарным.

Альтернативой пакету Statistica является, упомянутый ранее, интерпретируемый язык программирования и свободная программная среда, под одноименным названием “R” (Team et al., 2013). R поддерживает широкий спектр статистических

и численных методов и обладает хорошей расширяемостью с помощью пакетов – специализированных библиотек для работы специфических функций или специальных областей применения. Хотя в R используется интерфейс командной строки, доступны также и несколько графических интерфейсов пользователя. Как и в пакете Statistica, в R есть возможность создания качественной графики, с поддержкой математических символов.

Также стоит упомянуть распространенную среди экспериментаторов программу PAST (PAleontological STatistics) - комплексный, но простой в использовании программный пакет для выполнения ряда стандартных численных анализов и операций, используемых в количественной палеонтологии, и не только. PAST работает на стандартных компьютерах Windows и доступна бесплатно. Программа поддерживает ввод и представление данных в стиле электронных таблиц с одномерной и многомерной статистикой, реализует алгоритмы численной оптимизации, анализа временных рядов, построение графиков, и простого филогенетического анализа. Многие функции относятся к палеонтологии и экологии, и эти функции не встречаются в стандартных, более обширных статистических пакетах.

1.7 Заключение по обзору литературы и формулировка задачи исследования

Структура и форма колоса пшеницы являются важными характеристиками, непосредственно связанными с урожайностью и выживаемостью растений. Изучение генов, контролирующих данные признаки улучшит понимание генетических механизмов, для создания новых сортов с повышенной урожайностью и устойчивостью к различным факторам внешней среды. Сложность изучения генома пшеницы обусловлена в первую очередь полиплоидией, но в то же время это дает огромный потенциал в селекции. Изучение столь сложного генома пшеницы требует комплексного подхода. В современных сравнительных и селекционно-генетических исследованиях производится сбор и анализ данных о фенотипе и генотипе десятков тысяч растений. Такие условия требуют применения новых технологий для массового сбора, систематизации и анализа данных: для получения данных требуются современные методы фенотипирования растений; для сбора,

систематизации и хранения необходимо создание специализированных баз данных с использованием онтологий с целью унификации и стандартизации данных; для анализа требуется разработка моделей, описывающих данные экспериментов.

Разработке таких методов и их применению для фенотипирования пшеницы посвящены следующие разделы данной работы.

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1 Растительный материал

2.1.1 Растительный материал для морфометрии зерен

Для оценки морфометрических характеристик зерен были использованы пять генотипов пшеницы из коллекции зерновых культур лаборатории хромосомной инженерии злаков сектора генетики качества зерна Института Цитологии и Генетики СО РАН: Аленькая 1102 II-12, 84/98w 99 II-13, Synthetic 6x x-12, Purple Chance 4480 II-03 и Alcedo n-99 (таблица 2). Растения выращивались в поле в Новосибирске в 2014 году. Материал был выращен и предоставлен Пшеничниковой Т.А. Перечисленные сорта хорошо различаются по форме и размеру зерен. Указанные причины делают их удобными объектами для апробации разработанных методов.

Таблица 2. Размеры и форма зерен использованных сортов пшеницы.

Сорт	Форма	Средняя длина (мм)	Средняя ширина (мм)
Alcedo	овальная	7	3,6
Synthetic	удлиненная	8	2,3
Аленькая	-	5	2,4
84/98w	-	6,5	2,6
Purple Chance	-	7	2,9

2.1.2 Растительный материал для морфометрии колосьев

Для анализа колосьев были использованы 249 растений из коллекции д.б.н. Гончарова Н.П. Выборка включала растения восьми видов гексаплоидной пшеницы и F2 гибридов от скрещивания австралийского сорта мягкой пшеницы Triple Dirk с образцом KU506 китайской пшеницы *T. yunnanense* (син. *T. spelta* ssp. *yunnanense* (King ex S.L. Chen) N.P. Gontsch.). Название выбранных видов и гибридов, а также количество их образцов приведены в таблице 3.

Таблица 3. Виды и гибриды пшеницы, используемые для оценки точности разработанных методов. В строках таблицы представлено название вида, соответствующее число растений данного вида и число полученных фотографий колоса этих растений.

Вид	Количество растений	Количество фотографий
T. compactum	63	315
F2 Triple Dirk x Triticum yunnanense	52	260
T. aestivum	50	250

T. antiquorum	20	100
T. sphaerococcum	19	95
T. spelta	18	90
Amphyploid speltiforme	9	45
T. yunnanense	9	45
T. macha	9	45

T. compactum - вид пшеницы, гексаплоид с 21 хромосомой, адаптированный к условиям выращивания с низкой влажностью. *T. compactum* похож на мягкую пшеницу (*T. aestivum*), поэтому ее часто считают подвидом *T. aestivum compactum*. Его можно отличить по более компактному колосу из-за более коротких сегментов стержня, что дало ему общее название.

T. aestívum — пшеница мягкая, вид однолетних травянистых растений рода Пшеница (*Triticum*), характеризующийся плотными узкими колосьями с ломкой осью и снабженными длинными остями. Вид засухоустойчив, стоек к поражению ржавчиной и головнёй, не полегает. Колосья обычно пяти-цветковые. Относится к гексаплоидным пшеницам.

T. antiquorum – голозерная гексаплоидная пшеница, предшественница *T. compactum*.

T. sphaerococcum Perc. - пшеница шарозерная – узкоэндемичный вид гексаплоидной пшеницы, в прошлом распространенный на северо-западе Индии. Имеет яровой образ жизни, приспособлена к сухому климату в условиях орошения. Привлекла внимание селекционеров комплексом ценных признаков: шаровидная форма зерновки, устойчивость к полеганию, жаростойкость, неосыпаемость, высокие хлебопекарные качества. В то же время, пшеница шарозерная недостаточно устойчива к холодам, засухе и грибным болезням. Кроме того, ее урожайность существенно уступает стандартным сортам пшеницы мягкой.

Методом межвидовой гибридизации из пшеницы шарозерной и пшеницы мягкой озимой (сорт Обрий) в Краснодарском НИИСХ им. П.П. Лукьяненко был получен сорт пшеницы шарозерной озимой Шарада и производные от него сорта (Прасковья и Еремеевна), сочетающие в себе лучшие генетические качества исходных видов.

T. spelta — спельта относится к так называемой полбяной пшенице — группе видов с плёнчатым зерном и с ломкими колосьями. Выращивается с 5-го тысячелетия до нашей эры. Спельта является результатом естественной гибридизации пшеницы двузернянки (*T. dicoccum*) и дикорастущей пшеницы *(A. tauschii*). Эта гибридизация вероятно имела место на Ближнем Востоке, потому что именно здесь растет *A. tauschii* и это событие должно было произойти до появления обыкновенной пшеницы (*T. aestivum*). Генетические данные показывают, что спельта могла также возникнуть в результате гибридизации обыкновенной пшеницы и пшеницы двузернянки. Таким образом, гораздо более позднее появление полбы в Европе может быть результатом более поздней, второй гибридизации между пшеницы двузернянки и обыкновенной пшеницей.

T. yunnanense King ex S.L. Chen - гексаплоидный вид китайской пшеницы.

T. macha - подвид двузернянки (лат. *T. dicoccum*) однолетнего травянистого растения рода пшеницы, характеризующаяся плотными узкими колосьями с ломкой осью и снабженными длинными остями. Колосья обычно пяти-цветковые. Вид засухоустойчив, стоек к поражению ржавчиной и головнёй, не полегает.

Растения выращивали в гидропонной теплице при индивидуальной изоляции и стандартных условиях влажности, температуры и освещения. Колосья аннотировались вручную сотрудниками ИЦиГ СО РАН Фу Хао и Туманяном С.Р. Анализ структуры колоса (длина, ширина, остистость/безостость) проводили по стандартной методике. Плотность колоса определяли по формуле 1. Оценка остистости производилась только для F2 поколения межвидовых гибридов.

2.2 Методы анализа изображений

Методы анализа изображений были реализованы с использованием библиотеки OpenCV версии 3.4.3 (Dawson-Howe, 2014; Howse, 2015). Они включает в себя следующие этапы:

1. Предварительная обработка (подавление шумов и улучшение качества). Как правило, этот этап реализовывался на основе фильтра Гаусса (формула 2).

2. Сегментация изображения. Для сегментации изображения использовались как простые методы: адаптивная пороговая бинаризация полутоновых изображений, реализованная функцией adaptiveThreshold(…) библиотеки OpenCV, так и более специализированный подход: цветовая сегментация на основе диапазонов значений каналов цветовой модели HSV.

3. Выделение контуров. Поиск контуров осуществлялся с помощью функций findContours(…) библиотеки OpenCV, которая производит поиск замкнутых контуров на бинаризованном изображении. Также этот этап включает отбор интересующих контуров (исключение ложных объектов из обработки и др.).

4. Анализ формы объектов. Для полученных на предыдущем этапе контуров вычисляется ряд характеристик, таких как: площадь, периметр, индексы[1] формы контура, линейные размеры (длина, ширина). Для вычисления последних характеристик, необходимы дополнительная информация о структуре распознаваемых объектов (зерно или колос). Их вычисление будет описано далее в разделах 3.2.2 для зерен и в разделах 4.2.4, 4.2.6, 4.2.7, для колоса.

К общим характеристикам контура зерна и колоса можно отнести: моменты, центр масс, моменты инерции, главные оси, основная ось.

Моменты - это суммарные характеристики изображения:

$$M_{ij} = \sum_{x} \sum_{y} x^i y^j I(x, y)$$

где i, j – это индексы момента, x, y – координаты изображения, а I(x, y) – значение яркости пикселей изображения в точках (x, y). Моменты можно определить и для контура, если заменить сумму всех точек изображения на точки контура при условии, что для всех (x, y) точек контура I(n): I(x, y) = I(n) = 1.

Центр масс изображения (контура) - точка, соответствующая средневзвешенным координатам всех точек изображения (контура) $(\bar{x}, \bar{y})$, которую можно выразить через моменты изображения (контура): $\bar{x} = \frac{M_{10}}{M_{00}}$, $\bar{y} = \frac{M_{01}}{M_{00}}$.

[1] Индекс – числовая характеристика единичной размерности.

Момент инерции — ещё одна интегральная характеристика геометрической фигуры, пришедшая из физики и определяемая относительно некоторой оси вращения. По аналогии с моментами и центром масс момент инерции можно определить, как для изображения, так и для контура. Момент инерции изображения (контура) складывается из моментов инерции всех точек изображения (контура), относительно некоторой оси вращения l: $I_x = I(x,y)d^2$, где I(x,y) – значение яркости пикселей, а d – расстояние от точки (x, y) до оси вращения l. В случае контура I(x, y) = I(n) = 1.

Оси изображения (контура), момент инерции относительно которых равен нулю, называются главными (оси симметрии). Главные оси, проходящие через центр масс, называют главными центральными осями инерции. Произвольные же изображения (контуры) могут не иметь ни одной оси симметрии. Но ослабив критерий, можно определить тем самым более универсальную характеристику - основную центральную ось – прямую, проходящую через центр масс, момент инерции которой минимален. Общая блок-схема предложенных алгоритмов обработки изображения представлена на рисунке 12.

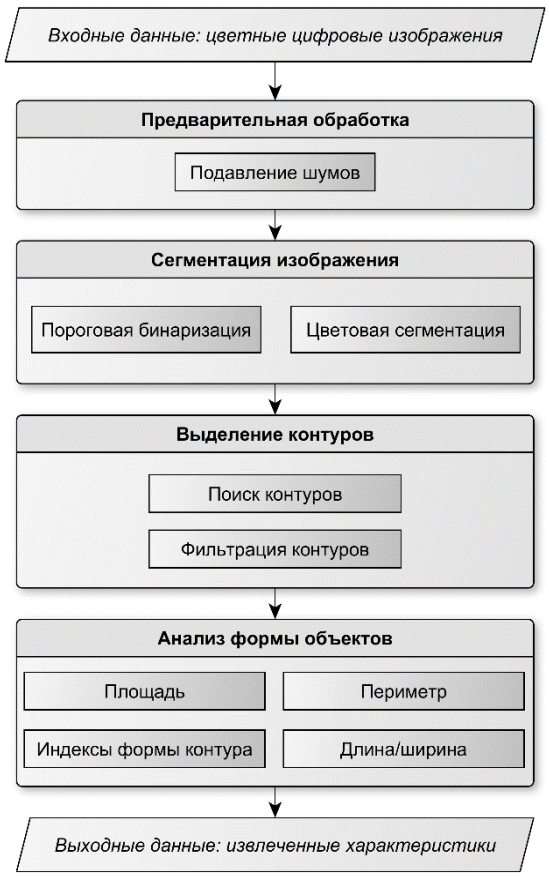


Рисунок 12. Блок-схема алгоритма обработки изображения. Ход алгоритма представлен сверху вниз. Параллелограммами представлены входные и выходные данные. Прямоугольные блоки представляют отдельные шаги алгоритма, объединенные в логические блоки-этапы.

2.3 Методы оценки точности алгоритмов анализа изображений

Оценка точности алгоритмов анализа изображений строится на основе двух типах оценок: оценка точности идентификации контуров распознаваемых объектов и оценки точности количественных характеристик, полученных на основе анализа изображения.

Точность автоматической идентификации контуров может быть сравнена с ручной идентификацией контуров на изображении. Для этого необходимо иметь набор предварительно размеченных вручную изображений с искомыми объектами. В качестве разметки при этом может выступать бинаризованный отпечаток распознаваемых объектов изображении, на котором пиксели интересующих объектов отмечены как 1, а фона - 0. Сопоставляя вручную размеченный отпечаток с аналогичной автоматизированной сегментацией можно посчитать оценку точности, принимая ручную сегментацию за эталонную. Для оценки точности алгоритма сегментации изображения на фон и колос использовался индекс Жаккара (Jaccard P., 1912):

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (3)$$

, где A – пиксели области изображения, полученной в результате сегментации с помощью разработанного алгоритма и заданных значений его параметров. B – пиксели области изображения, маркированные вручную. Индекс варьируется от 0 - полное несовпадение, до 1 - полное совпадение (рисунок 13).

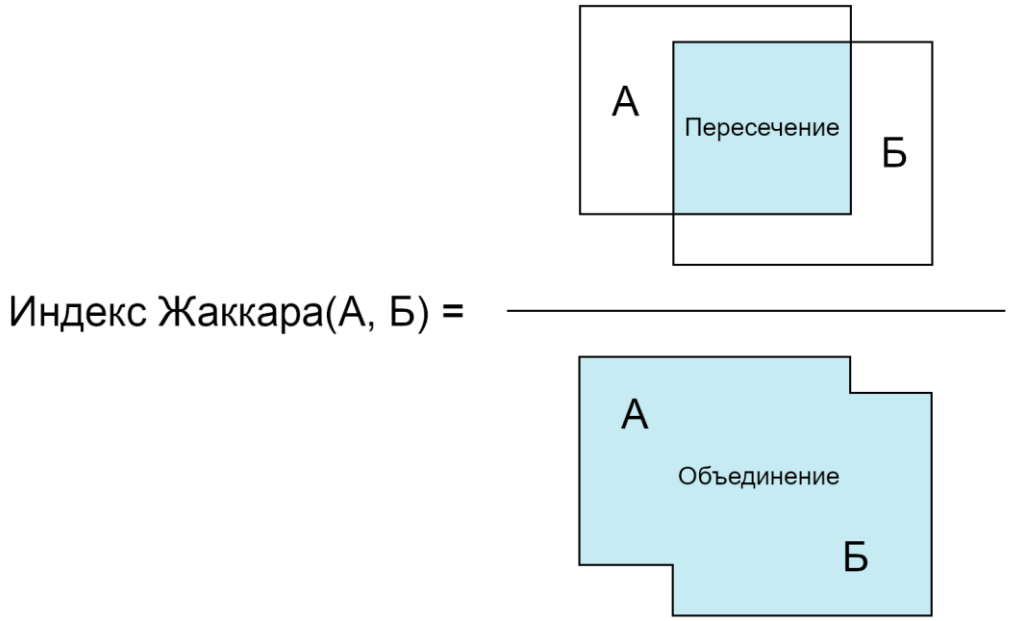


Рисунок 13. Индекс Жаккара может быть использован для сравнения совпадения областей на изображении. На рисунке индекс Жаккара схематически представлен дробью. Буквами А и Б обозначены сравниваемые области. В верхней части представлены границы областей на изображении и их пересечение (закрашенная область). В нижней части представлено объединение данных областей (закрашенная область). Индекс Жаккара равен отношению числа пикселей в пересечении областей к числу пикселей в их объединении.

Точность количественных характеристик, полученных на основе анализа изображений, определяется путем их сравнения с характеристиками, измеренными при ручном фенотипировании. Для сравнения наборов скалярных значений подходят средняя абсолютная ошибка (MAE, mean absolute error) и средняя абсолютная процентная погрешность (MAPE, mean absolute percent error) вычисляемые следующим образом:

$$MAE = \frac{1}{M}\sum_{j=1}^{j=M}\left|n_j - {n'}_j\right| \quad (4)$$

$$MAPE = \frac{100}{M}\sum_{j=1}^{j=M}\left(\frac{\left|n_j - {n'}_j\right|}{n_j}\right) \quad (5)$$

где j - номер измерения, n_j – значение, полученное при фенотипировании вручную, n_j' – значение, полученное на основе анализа изображения, а M - количество измерений.

Чем больше значения MAE (4) и MAPE (5) тем больше значение ошибки измерений компьютерным методом. Если значения MAE и MAPE близки к 0, то и ошибка низкая.

Дополнительно, был оценены коэффициенты корреляции Пирсона, Кендалла и Спирмена (R_n, K_n, S_n) между значениями n_j и n_j. Чем ближе их значения к единице, тем меньше погрешность оценки.

2.4 Методы статистического анализа

Статистический анализ проводился с помощью свободного программного обеспечения: программной среды R и программного пакета для анализа научных данных PAST.

С помощью пакета статистического анализа R выполнялись:

- Дисперсионный анализ (однофакторный, двухфакторный). Дисперсионный ANOVA (ANalysis Of VAriance) позволяет оценивать влияние различных факторов на экспериментальных данных путём исследования значимости различий в средних значениях, учитывая степень их отклонения от средних;
- Рассчитывались коэффициенты корреляции;

С помощью пакета PAST выполнись:

- Кластерный анализ характеристик формы колоса;
- Анализ методом главных компонент. Метод главных компонент позволяет определять характеристики, вносящие наибольший вклад в дисперсию исходных данных;
- Рассчитывались таблицы парных различий характеристик колоса по критерию Манна-Уитни с поправкой Бонферрони на множественное сравнение гипотез.

2.5 Разработка приложения для Android

Для реализации мобильного приложения SeedCounter использовалась интегрированная среда разработки для работы с платформой Android - AndroidStudio. Обработка изображений на платформе Android осуществлялась с помощью специальной сборки библиотеки OpenCV для мобильных устройств. Приложению требуется устройство с Android 4.x, версии API 15 или выше, и доступ к камере устройства.

2.6 Методы реализации баз данных

Современная база данных как правило включает в себя следующий стек программного обеспечения: система управления базами данных (СУБД), web-сервер и web-интерфейс.

В качестве хранилища данных использовалась реляционная база данных под управлением СУБД MySQL, развернутая на сервере под управлением CentOS Linux. Реляционные БД основаны на развитом математическом аппарате и обладают такими преимуществами как полная независимость данных и минимальная избыточность (нормализация) данных. В качестве преимуществ MySQL стоит отметить широкую гибкость этой СУБД, высокую производительность и безопасность.

Web-сервер обеспечивает доступ к базе данных посредством сети Интернет, по протоколам HTTP/HTTPS в рамках клиент-серверной архитектуры. Клиентом в данной архитектуре выступает web-браузер. В качестве web-сервера был использован Nginx, как простой, быстрый и надёжный сервер, не перегруженный функциями.

Web-интерфейс был реализован на основе каркаса веб-приложений (CMF, Content Management Framework) и системы управления содержимым (CMS, Content Management System) Drupal (Abbott and Jones, 2016). Drupal написана на языке PHP и является свободным программным обеспечением (распространяется под лицензией GNU GPL 2+). Drupal поддерживает важный стандартный функционал, такой как: управление контентом, регистрация пользователей, управление правами пользователя и доступом к ресурсам, задает MVC модель системы и т.п. Один из

ключевых принципов Drupal это модульная структура с возможностью модифицировать поведение отдельных его частей без непосредственного вмешательства в уже имеющийся код, что позволяет Drupal оставаться достаточно гибким и эффективным инструментом для web-разработки. Для генерации конечных HTML-страниц Drupal 8 использует шаблонизатор Symphony. Шаблонизатор способствует корректному отделению представление данных от исполняемого кода согласно модели MVC, что повышает поддерживаемость и гибкость разработанного функционала.

ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1 Метод морфометрии зерен пшеницы с помощью мобильных устройств

3.1.1 Протокол получения изображения зерен пшеницы

Важный аспект получения изображений - это освещение. Оно может быть, как естественным, так и искусственным. К естественному освещению можно отнести дневной свет: прямой и отраженный свет внутри и вне помещений, в солнечную и в пасмурную погоду, который можно охарактеризовать освещенностью в люксах (лк). Искусственное же освещение создается при помощи осветительных приборов и характеризуется излучаемым светопотоком, цветовой температурой, частотой мерцания и зависит от числа и характеристик осветительных приборов. Важно также направление и расстояние источников света относительно освещаемых предметов. В рамках предложенного протокола, для снижения эффектов теней от зерен, необходимо минимум два источника искусственного освещения расположенных друг на против друга на расстоянии 20-40 см и направленных вниз, на снимаемую поверхность. В случае естественного освещения необходимо обеспечить максимально возможное освещение рассеянными лучами. Прямые солнечные лучи могут усиливать искажения, связанные с тенями от зерен. С учетом этих особенностей нами был разработан оригинальный протокол фенотипирования зерен пшеницы при помощи мобильных устройств.

При подготовке к измерению зерна из колоса очищаются от чешуи и мусора, и рассыпаются на белый лист бумаги формата A4, Letter, Legal, A3, A5, B4, B5 или B6 произвольным образом. Лист при этом размещается на темной поверхности, контрастной по отношению к листу, для упрощения его последующего распознавания.

Для уменьшения ошибок следует обеспечить следующие условия:

- направление света должно быть сверху-вниз под максимально прямым углом по отношению к поверхности листа бумаги, насколько это возможно;

- устройство получения изображений (фотокамера или мобильное устройство) располагается на расстоянии около 50 см перпендикулярно над листом.
- границы листа бумаги и фона должны быть максимально параллельны сторонам фотокадра (рисунок 14).

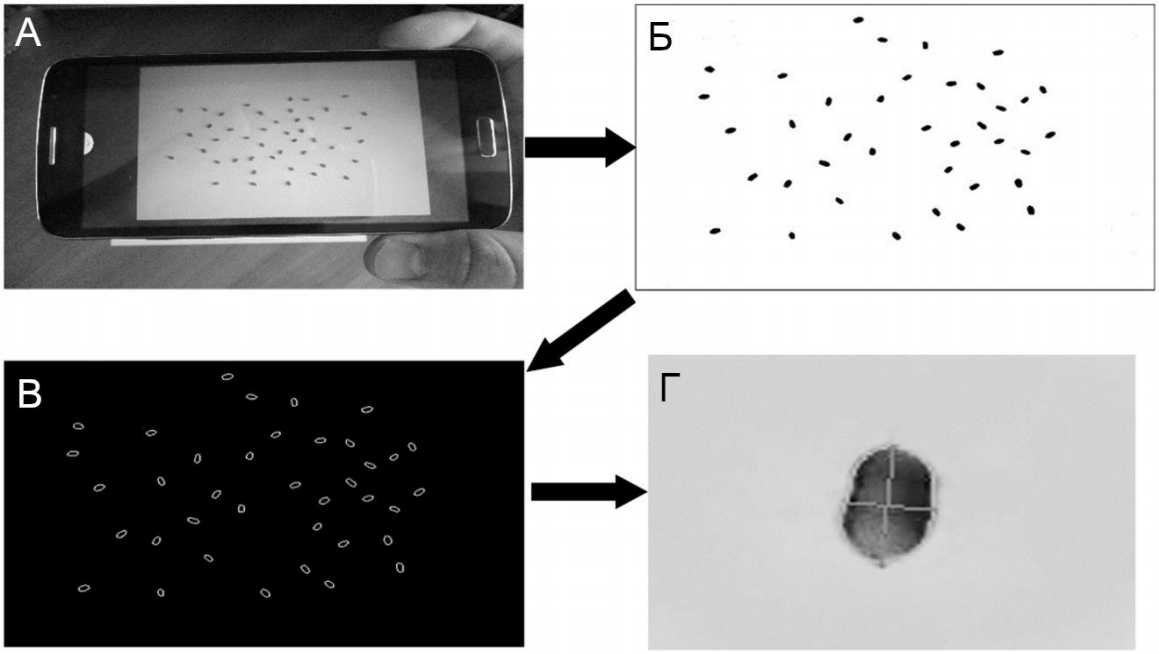


Рисунок 14. Основные этапы распознавания зерна на листе бумаги. Стрелками указан порядок следования этапов. (А) Захват изображения с помощью камеры мобильного устройства и распознавания листа бумаги. (Б) Изображение после аффинного преобразования и бинаризации. (В) Контуры зерна, идентифицированные на изображении. (Г) Изображение зерна с основными осями, показанное скрещенными линиями.

Фиксированный размер бумаги позволяет рассчитать масштаб изображения и оценить размеры зерен в метрических единицах.

3.2 Алгоритм анализа изображений для определения характерных черт зерен пшеницы

Алгоритм реализован с использованием библиотеки обработки изображений OpenCV (Howse, 2015; Dawson-Howe, 2014) и состоит из нескольких этапов:

- распознавание листа бумаги и аффинные преобразования для коррекции перспективы;

- распознавание и фильтрация контуров зерен;
- морфометрия распознанных зерен.

Рассмотрим каждый этап более детально.

3.2.1 Распознавание листа бумаги

Лист бумаги распознается как светлая область четырехугольной формы на темном фоне. Для выполнения коррекции перспективы на следующем этапе необходимо вычислить координаты углов листа бумаги на исходном изображении. Для достижения этого выполняется следующее:

- Преобразование исходного цветного изображения в оттенки серого. Изображение в градациях серого необходимо для последующей бинаризации по пороговой интенсивности яркости пикселей. Преобразование вычисляет среднее значение интенсивности по каждой из компонент каналов RGB для каждого пикселя изображения отдельно (с помощью функции cvtColor(…)).

- Пороговая бинаризация - позволяет получить черно-белое изображение с явно определенными областями фона и листа бумаги. Используется адаптивный алгоритм для уменьшения эффектов засвечивания и затемнения на изображении (функция adaptiveThreshold(…)). Алгоритм определяет порог для пикселя на основе небольшой области вокруг него. Таким образом, для разных областей изображения применяются разные пороговые значения, что дает лучшие результаты для изображений с разным освещением.

- Распознавание границ листа – требуется для последующей аппроксимации прямыми линиями. Оператор Кэнни используется для поиска всех границ изображении (функция canny(…)). На выходе получается изображение с белыми границами на черном фоне.

- Поиск границ листа с помощью преобразования Хафа (функции houghLinesP(…)). Параметры алгоритма определяют длину аппроксимирующих отрезков от 20% до 80% относительно ширины

изображения. По причине различных искажений, в основном возникающих на предыдущих этапах, полученное множество может содержать отрезки некорректно аппроксимирующие границы. К примеру: отрезки, не относящиеся к границам листа (найденные границы на фоне или внутри области листа); короткие отрезки, лежащие вдоль части границы (разрыв границы по причине блика); отрезки, лежащие вдоль границы со смещенным углом относительно границы листа. Поэтому, для более точного распознавания границ листа, алгоритм кластеризует полученные отрезки на основе их геометрического расположения, формируя четыре кластера, которые соответствуют сторонам листа. Для каждого кластера восстанавливается граница листа, объединяя отрезки внутри него: усредняя пересекающиеся отрезки, и сращивая концы непересекающихся.

- Вычисление координат углов листа – определяются путем вычисления точек пересечений распознанных граничных линий листа.

Если форма листа бумаги на изображении отличается от прямоугольной, алгоритм трансформирует исходное изображение с помощью аффинных преобразований, чтобы она стала прямоугольной путем аффинных преобразований. Этот шаг выполняется при помощи функции getPerspectiveTransform(…) для вычислений матрицы преобразования и функции warpPerspective(…) для преобразования изображения. На итоговом изображении противоположные ребра распознанных границ листа бумаги становятся параллельными, все углы равны 90º, а области за границами листа исключаются из дальнейшей обработки.

3.2.2 Идентификация и морфометрия зерен

Чтобы распознать контуры зерен на изображении со скорректированной перспективой, вновь производится пороговая бинаризация с порогом, разделяющим зерна на листе бумаги. После чего, используется алгоритм поиска замкнутых контуров, применяя функцию findContours(…) к бинаризованному изображению. Функция возвращает список найденных контуров, представленных в виде упорядоченного множества точек.

При определении границ зерен для уменьшения влияния затенения производится их корректировка. Локальная прямоугольная область, содержащая идентифицируемый контур зерен, копируется в отдельное изображение такого же размера, преобразовывается в цветовое пространство HSV и сегментируется. Локальная бинаризация дает более точные определения границ зерен по сравнению с бинаризацией всего изображения при поиске контуров.

Одна из трудностей в определении границ зерна является их соприкосновение на изображении. Для установления границ между зернами, которые соприкасаются на изображении используется метод водораздела с маркерами (Roerdink and Meijster, 2000) (функция watershed(…)).

Маркеры были получены следующим образом: сначала анализировался контур слипшихся зерен, чтобы определить, сколько зерен потенциально объединены в одном контуре. Для этого вычислялись точки перегибов контура, направления и попарные соответствия перегибов контура. На основании полученных данных оценивалось число зерен в едином контуре. Для каждого отдельного контура слипшихся зерен отрисовывался отпечаток на отдельном пустом изображении. Далее последовательно выполнялось дистанционное (distanceTrainsform(…)) и пороговое преобразование изображения. Дистанционное преобразование заполняет области отпечатков контуров значениями, равными расстоянию до ближайшей точки контура (ближайшей точки фона). Пороговое преобразование в таком случае “отрезает” переходные области между зернами – там, где проходит их граница. Остаются только гарантированные области зерен – наиболее удаленные от точек контура пиксели. Полученные в результате отпечатки области зерен определялись как маркеры.

Полученные контуры зерен аппроксимируются эллипсоидами, позволяя оценивать размер главных осей, соответствующих длине и ширине зерна (рисунок 14 Г).

3.3 Мобильное приложение SeedCounter

Данный алгоритм реализован в мобильном и настольном приложении SeedCounter. Мобильная версия SeedCounter – это приложение для платформы

Android, которое получает изображения непосредственно с камеры мобильного устройства и выполняет автоматический расчет морфологических параметров зерна пшеницы с помощью мобильных устройств в полевых условиях, без необходимости использования ноутбуков или персональных компьютеров. Приложение позволяет распознать зерна на изображении листа бумаги формата Letter, Legal, A3, A4, A5, B4, B5 или B6, оценить их количество и также морфометрические параметры, такие как длина, ширина, проецируемая на бумажный лист площадь и расстояние между геометрическим центром массы зерна и точки пересечения его главных осей. Проведено несколько тестов подсчета зерен в условиях контролируемого освещения и дневного света для оценки производительности программного обеспечения.

3.3.1 Интерфейс мобильного приложения

Пользователь мобильного приложения может настраивать параметры обработки изображений и распознавания зерен с помощью опции «Калибровка» в главном меню (рисунок 15 А).

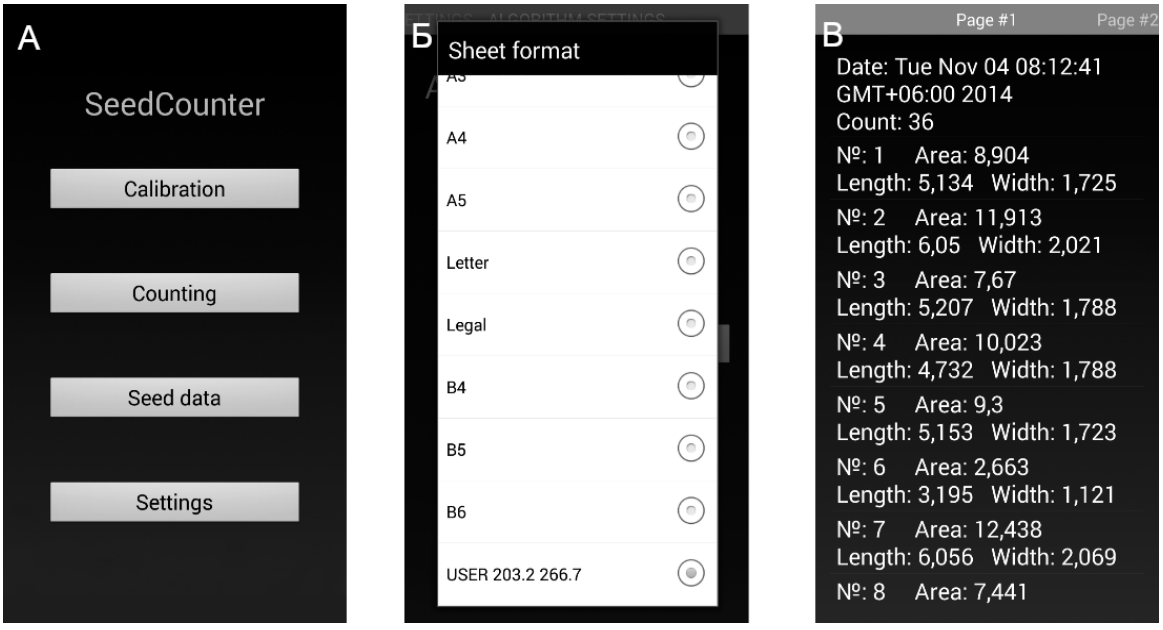


Рисунок 15. Интерфейс приложения SeedCounter. Панели слева-направо: А) Главное меню. Позволяет перейти к одному из действий: калибровке, подсчету зерен, просмотру сохраненных данных, переходу в меню опций. Б) Меню выбора формата листа бумаги. В) Экран вывода результатов измерений (количество зерен и длина/ширина/площадь каждого зерна).

Для этого пользователь должен разместить одно зерно на бумаге и убедиться, что алгоритм правильно распознает зерно и отмечает его в красном многоугольнике. Параметры алгоритма на этом этапе сохраняются автоматически. Пользователь может также использовать меню программы (рисунок 15 Б), чтобы задать размер листа бумаги (включая определяемые пользователем размеры); разрешение камеры. Также интерфейс позволяет включить/отключить алгоритм разделения зерен и уточнение границ зерна. Данные об общем числе зерен и количественных характеристиках каждого отдельного зерна отображаются в меню «Seed data» (рисунок 15 В). Информация может быть экспортирована в формат широко используемого расширяемого языка разметки (eXtensible Markup Language, XML) или текстового формата представления таблиц, (Tab Separated Values, TSV). Данные можно отправить на веб-сервер в систему WheatPGE. В результате пользователь получает статический URL-адрес сгенерированной web-страницы с этими данными.

3.4 Оценка точности SeedCounter

Для оценки точности приложения SeedCounter были рассмотрены два типа ошибок. Во-первых, оценивалась точность идентификации числа зерен. Пятьдесят зерен пшеницы одного сорта высыпали на бумажный лист, а количество зерен оценивалось с помощью приложения SeedCounter. После этого с листа удаляли одно зерно, а остальные зерна перетасовывали (без ручного разделения зерен). После этого количество зерен оценивалось снова. Эту процедура повторялась 40 раз. Серии измерений (40 итераций) проводились при различных условиях освещения и на различных мобильных устройствах.

Во-вторых, оценивалась точность измерения длины и ширины отдельных зерен. Характеристики измерялись для 250 зерен пяти сортов пшеницы на мобильном устройстве и микроскопе. Были рассчитаны MAE и MAPE (см. формулы 4 и 5) измерений длины и ширины каждого зерна. Для них же были рассчитаны коэффициенты корреляции Пирсона (R), Кендалла (K) и Спирмена (S).

В-третьих, чтобы определить область применения морфометрии зерен с помощью мобильных устройств требовалось сопоставить оценки точности, разработанного метода, с существующими на данный момент подходами. Оценки

точности измерений характеристик размеров зерен пшеницы SeedCounter были сравнены с SmartGrain, ближайшем аналогом, имеющим заведомо больший потенциал точности благодаря большему разрешению изображений и фиксированным условиям их получения за счет использования планшетного сканера (Tanabata et al., 2012). Для этого была получена серия изображений с разрешением 600 dpi на сканере HP Scanjet 3800. Предполагалось, что иные подходы, использующие планшетные сканеры, такие как GrainScan, имеют аналогичную точность (Whan et al., 2014).

В экспериментах использовались три мобильных устройства под управлением ОС Android при максимальных разрешениях их камер: смартфоны Samsung Galaxy Grand 2, Sony Ericsson XPERIA pro mini и планшет DNS AirTab m101w. Характеристики этих устройств представлены в таблице 4.

Таблица 4. Характеристики использованных мобильных устройств и разрешения их камер.

Мобильное устройство	**Операционная система**	**Процессор (количество ядер x частота)**	**RAM**	**Разрешение камеры**
Sony Ericsson Experia pro mini	Android 2.3	Qualcomm MSM 8255 (1 x 1000 МГц)	512MB	5 Mpx (2592x1944)
DNS AirTab m101w	Android 4.1	RockChip RK3066 (2 x 1500 МГц)	1GB	5 Mpx (2592x1944)
Samsung Galaxy Grand 2	Android 4.3	ARM Cortex-A7 (4 x 1200 МГц)	1.5GB	8 Mpx (3264x2448)

В качестве осветительных устройств применялись: лампа дневного света 11 Вт (цветовая температура 4000 К, световой поток 900 лм), лампа дневного света 5 Вт (4000 К, 400 лм) и галогенная лампа 35 Вт (2700 К, 190 мкм). Они использовались для формирования четырех конфигураций искусственного освещения: лампа дневного света 11 Вт (L1); лампа дневного света 11 Вт и две лампы дневного света мощностью по 5 Вт (L2); лампа дневного света 11 Вт и четыре лампы дневного света по 5 Вт (L3); и лампа дневного света 11 Вт, четыре лампы дневного света по 5 Вт и галогеновая лампа 35 Вт (L4). Лампы были установлены на высоте 60 см над листом бумаги. Лист размещался на столе с темным покрытием. Измерения проводились в темной комнате. Чтобы оценить точность измерений в дневное время, были измерены зерна без использования искусственного освещения в облачную погоду в закрытом помещении и в ясный день на открытом воздухе.

В естественных условиях освещения для измерения светового потока была использована цифровая зеркальная фотокамера. При этом оценивалось значение экспозиции (exposure value, EV), при светочувствительности ISO 100 и при фиксированных значениях выдержки и диафрагмы. После чего определялось соответствующее значение освещенности в люксах (лк).

Подробная информация об условиях эксперимента приведена в таблице 5.

Таблица 5. Условия освещения для измерения точности приложения SeedCounter.

Условное обозначение	**Условия освещения**	**Световой поток (люмен, люкс)**	**Температура света**
L1	Лампа дневного света, 11 ватт	900 лм	4000K
L2	Лампа дневного света, 11 ватт и две лампы дневного света по 5 ватт	1700 лм	4000K
L3	Лампа дневного света, 11 ватт и четыре лампы дневного света по 5 ватт	2500 лм	4000K
L4	Лампа дневного света, 11 ватт, четыре лампы дневного света по 5 ватт и галогеновая лампа, 35 ватт	2690 лм	4000K и 2700K
L5	Дневной свет, снаружи, пасмурная погода	(1280 лк)	–
L6	Дневной свет, снаружи, солнечная погода	(656000 лк)	–

Двусторонние тесты ANOVA были использованы для оценки влияния типа устройства и условий освещения на точность подсчета количества зерен, их длины и ширины. Тип устройства и освещение рассматривались как независимые переменные, а оценки ошибок (MAE и MAPE) как зависимые переменные. Для выполнения этого теста использовалось программное обеспечение Statistica 6.0 (Халафян, 2010).

Точность оценки количества зерен для разных серий экспериментов показана в таблице 6.

Таблица 6. Оценка точности подсчета количества зерен приложением SeedCounter. MAE – средняя абсолютная ошибка числа зерен, MAPE – средняя абсолютная относительная ошибка числа зерен, $r(N_i, N_j')$ – коэффициент корреляции Пирсона числа зерен посчитанных с помощью приложения SeedCounter и вручную.

Условия освещения	**Тип устройства**	**MAE (шт)**	**MAPE (%)**	r **(**N_j**,** N_j**')**
L1	Samsung Galaxy Grand 2	1,425	3,5	0,996
L2	Samsung Galaxy Grand 2	1,375	3,6	0,994
L3	Samsung Galaxy Grand 2	0,65	1,5	0,998
L4	Samsung Galaxy Grand 2	0,975	2,4	0,997
L5	Samsung Galaxy Grand 2	1,15	2,9	0,992
L6	Samsung Galaxy Grand 2	0,55	1,7	0,998
L1	Sony Ericsson Experia pro mini	1	2,4	0,995
L2	Sony Ericsson Experia pro mini	0,8	1,9	0,995
L3	Sony Ericsson Experia pro mini	0,675	1,7	0,996
L4	Sony Ericsson Experia pro mini	0,775	2,0	0,997
L5	Sony Ericsson Experia pro mini	0,75	1,8	0,996
L6	Sony Ericsson Experia pro mini	0,775	1,8	0,996
L1	DNS AirTab m101w	1,2	3,1	0,997
L2	DNS AirTab m101w	0,5	1,2	0,997
L3	DNS AirTab m101w	0,125	0,3	0,999
L4	DNS AirTab m101w	0,725	1,7	0,998
L5	DNS AirTab m101w	1,175	3,0	0,996
L6	DNS AirTab m101w	0,775	2,0	0,997

Таблица показывает, что средняя относительная ошибка подсчета числа зерен не высока, и составляет 2% относительно общего числа зерен, а абсолютная ошибка в среднем равна 1 зерну.

Более подробный анализ показал, что ошибки при подсчете количества зерен чаще всего возникают из-за близкого размещения зерен на листе, когда два (или более) зерна соприкасаются и распознаются как единый объект. Данная проблема усугубляется в случае плохого освещения. При устранении соприкасающихся зерен, ошибка в подсчете их числа становится равной 0.

Далее была проведена оценка точности определения размеров зерен при помощи разработанного приложения. Точность оценки длины и ширины зерен мобильными устройствами в разных условиях освещения приведена в таблице 7.

Таблица 7. Точность оценки длины (верхняя ячейка) и ширины (нижняя ячейка) зерен пшеницы измеренных с помощью приложений SeedCounter и SmartGrain. Приведены значения: MAE – средняя абсолютная ошибка; MAPE – средняя абсолютная относительная ошибка; R, K, S – коэффициенты корреляции Пирсона, Кендалла и Спирмена, соответственно; для длины/ширины зерен, оцененных с помощью приложений SeedCounter/SmartGrain, и измеренных вручную, с помощью приложения ImageJ, на изображениях с микроскопа.

			Оценка по длине/ширине				
Программный продукт	**Условия освещения**	**Тип устройства**	**MAE (mm)**	**MAPE (%)**	***R***	**K**	**S**
SeedCounter	L1	Samsung Galaxy Grand 2	0,287	4,237	0,936	0,758	0,912
			0,281	10,669	0,816	0,513	0,695
		Sony Ericsson Experia pro mini	0,313	4,612	0,933	0,749	0,909
			0,312	11,941	0,765	0,422	0,584
		DNS AirTab m101w	0,295	4,350	0,943	0,769	0,922
			0,296	11,355	0,774	0,446	0,611
SeedCounter	L2	Samsung Galaxy Grand 2	0,311	4,573	0,928	0,747	0,905
			0,281	10,578	0,825	0,521	0,707
		Sony Ericsson Experia pro mini	0,317	4,664	0,931	0,748	0,906
			0,303	11,578	0,767	0,429	0,593
		DNS AirTab m101w	0,306	4,476	0,935	0,759	0,913
			0,286	10,900	0,777	0,456	0,627
SeedCounter	L3	Samsung Galaxy Grand 2	0,291	4,313	0,932	0,755	0,911
			0,276	10,365	0,822	0,519	0,704
		Sony Ericsson Experia pro mini	0,306	4,475	0,937	0,757	0,912
			0,289	11,099	0,777	0,449	0,613
		DNS AirTab m101w	0,284	4,166	0,950	0,780	0,928
			0,289	11,294	0,779	0,451	0,618
SeedCounter	L4	Samsung Galaxy Grand 2	0,355	5,222	0,923	0,738	0,902
			0,298	11,390	0,811	0,496	0,675
		Sony Ericsson Experia pro mini	0,349	5,115	0,920	0,731	0,897
			0,306	11,722	0,755	0,426	0,586
		DNS AirTab m101w	0,318	4,684	0,940	0,768	0,922
			0,304	11,775	0,780	0,448	0,615
SeedCounter	L5		0,491	7,341	0,797	0,593	0,766

		Samsung Galaxy Grand 2	0,304	10,820	0,770	0,507	0,685
		Sony Ericsson Experia pro mini	0,319	4,742	0,913	0,737	0,900
			0,283	10,712	0,750	0,468	0,635
		DNS AirTab m101w	0,393	5,923	0,890	0,684	0,861
			0,309	11,674	0,672	0,413	0,563
SeedCounter	L6	Samsung Galaxy Grand 2	0,409	6,233	0,875	0,658	0,843
			0,289	10,641	0,769	0,481	0,656
		Sony Ericsson Experia pro mini	0,362	5,566	0,899	0,693	0,871
			0,313	11,527	0,730	0,438	0,595
		DNS AirTab m101w	0,416	6,319	0,890	0,689	0,865
			0,276	10,209	0,787	0,510	0,695
SmartGrain	-	HP Scanjet 3800	0,391	5,617	0,948	0,778	0,926
			0,219	8,328	0,886	0,674	0,854

Таблица показывает, что ошибка оценки размера зерна была равна приблизительно 0,30 мм (средняя для всех серий: 0,31 мм), что составляет около 8% от линейных размеров зерна (средний для всех серий: 8,03%). Коэффициенты корреляции между реальным размером длины зерен (измеренное вручную) и её программной оценкой во всех экспериментах были не ниже 0,79 (коэффициент корреляции Пирсона, $p<0,01$). Высокие значения коэффициентов корреляции наблюдаются между реальным размером ширины зерен и ее программной оценкой. В этом случае коэффициент Пирсона больше 0,67 во всех экспериментах ($p<0,01$). Интересно, что ошибки оценок длины зерна для SeedCounter и SmartGrain близки друг к другу. Однако для ширины зерен SmartGrain демонстрирует большую точность, вероятно, в связи с меньшей подверженностью теневым искажениям. Стоит отметить, что SmartGrain анализировал изображения, полученные со сканера, в фиксированных условиях освещения и с высоким разрешением. SeedCounter же в свою очередь обрабатывает изображения, разрешение которых заведомо ниже, полученных в различных условиях освещения, зачастую далеких от идеальных. Эти факторы заведомо ограничивают потенциальную точность приложения SeedCounter.

В таблице 8 указаны средние значения измерений с помощью, разработанной программы при различных условиях освещения.

Таблица 8. Средние значения точности оценки приложением SeedCounter длины и ширины в различных условиях освещения. MAE – средняя абсолютная ошибка длины и ширины зерна. MAPE – средняя абсолютная относительная ошибка длины и ширины зерна. R, K, S – средние значения коэффициентов корреляции Пирсона, Кенделла и Спирмана для длины и ширины зерна, оцененной SeedCounter и измеренной вручную.

Условия освещения	**Среднее MAE**	**Среднее MAPE**	**Среднее R**	**Среднее K**	**Среднее S**
L1	0.297	7.861	0.861	0.610	0.772
L2	0.301	7.795	0.860	0.610	0.775
L3	0.289	7.619	0.866	0.618	0.781
L4	0.322	8.318	0.855	0.601	0.766
L5	0.350	8.535	0.799	0.567	0.735
L6	0.344	8.416	0.825	0.578	0.754
SmartGrain	0.305	6.973	0.917	0.726	0.890

Мобильные устройства в среднем демонстрируют наилучшую точность при оценке размера зерен в условиях освещения двух ламп дневного света и светового потока в размере 2500 люмен (условие освещения L3). Наихудшая точность была достигнута в помещении в пасмурный день (условие освещения L5). Двухсторонний тест ANOVA показал, что условия освещения существенно влияют на точность измерений программным продуктом (ANOVA, p<0,05). Данные вычислений приведены в таблице 9.

Таблица 9. Значимость влияния типа мобильного устройства и условий освещения на ошибку подсчета количества зерен и оценки их размеров. Двухфакторный дисперсионный анализ ANOVA (p-values). Жирным отмечены значимые значения (p<0.05).

Тип ошибки	**Условия освещения**	**Тип устройства**
Подсчет зерен, MAE	**0,004**	0,365
Подсчет зерен, MAPE	**0,003**	0,306
Размеры зерен, MAE	**0,036**	0,771
Размеры зерен, MAPE	0,094	0,890

Стоит отметить, что наибольшее среднее значение MAE при подсчете количества зерен (среднее значение равно 0,458) получено для освещения с самым низким световым потоком L1 (одна лампа, 11 Вт). При других источников освещения наблюдаются более низкие значения: 0,058 для L2; 0,1 для L3; 0,058 для L4 и 0,275 для L5. Также, следует отметить, что MAPE подсчета зерен в условиях без искусственного света меньше, чем для самого низкого светового потока, но больше, чем при всех других условиях освещения. Это может быть объясняться тем, что одиночный источник искусственного света, способен усиливать затенение, что не лучшим образом сказывается на точности распознавания границ объектов.

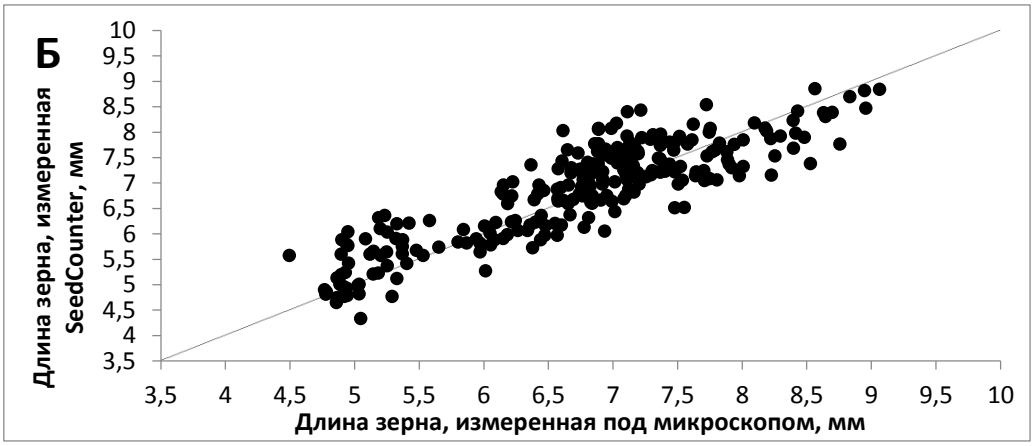
Результаты показывают, что тип устройства не оказывает существенного влияния на измерения количества зерен и их размеров. На рисунке 16 показана диаграмма рассеяния для длины и ширины 250 семян на изображениях, полученных с помощью микроскопа и мобильного устройства Samsung при дневном свете в пасмурную (L3) и солнечную (L6) погоду.

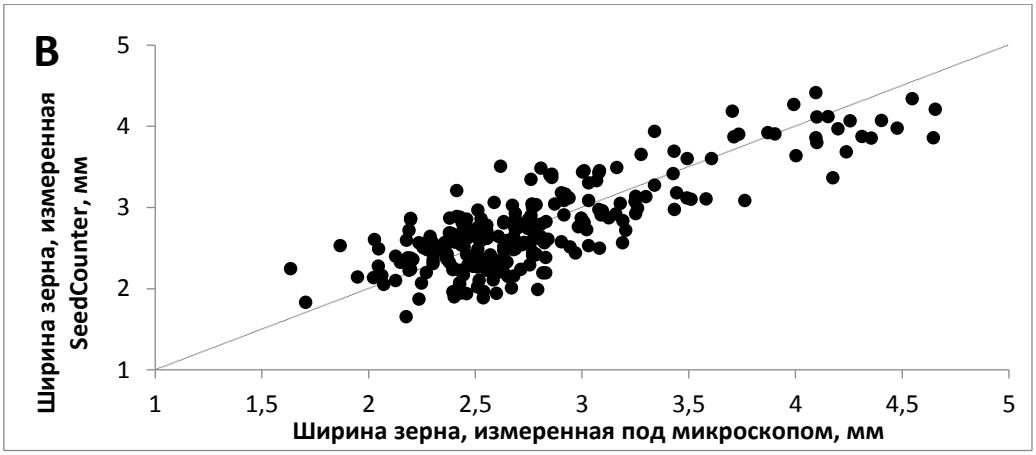
Данный рисунок демонстрирует, что при хороших условиях освещения оценки размера зерна, полученные мобильным устройством, согласуются с измерениями, полученными вручную, с помощью микроскопа. Однако в условиях солнечного света наше программное обеспечение имеет тенденцию недооценивать размеры зерен для крупных зерен и переоценивать их для меньших зерен. Этот эффект, вероятно, связан с затененностью, который вводит систематическое смещение в оценке размера зерна, когда изображение берется под прямым ярким солнечным светом.

Используя мобильное приложения SeedCounter была выполнена морфометрия зерен пшеницы пяти сортов. Для каждого сорта анализировались 50 зерен и измерялись их длина и ширина. Диаграммы на рисунке 17 A-B демонстрируют надежность дифференциации зерен из разных сортов пшеницы на основе их оценок длины и ширины. На рисунке показано, что сорт *Alcedo* имеет самые толстые зерна (средняя ширина - 3,59 мм), а сорт *Synthetic* имеет самые длинные зерна (7,97 мм), что соответствует их реальным размерам. Разделение сортов по размеру зерна четко показано на рисунке 17 B, где различные сорта занимают разные участки. Это показывает, что разработанное приложение можно использовать для идентификации сортов пшеницы с характерными, отличительными размерами зерна.

Было оценено время, используемое для анализа одного изображения мобильными устройствами и программным обеспечением SmartGrain при различных разрешениях изображения. Время обработки изображений с низким разрешением (Sony, 2592 × 1944 пикселей) составляет приблизительно 30 с. Для камеры с более высоким разрешением (Samsung, 3264 × 2448) это значение близко к 1 мин. Интересно, что это сопоставимо со временем обработки изображений SmartGrain (при аналогичных разрешениях 3510 × 2550).


А
Длина зерна, измеренная SeedCounter, мм
Длина зерна, измеренная под микроскопом, мм


Б
Длина зерна, измеренная SeedCounter, мм
Длина зерна, измеренная под микроскопом, мм


В
Ширина зерна, измеренная SeedCounter, мм
Ширина зерна, измеренная под микроскопом, мм

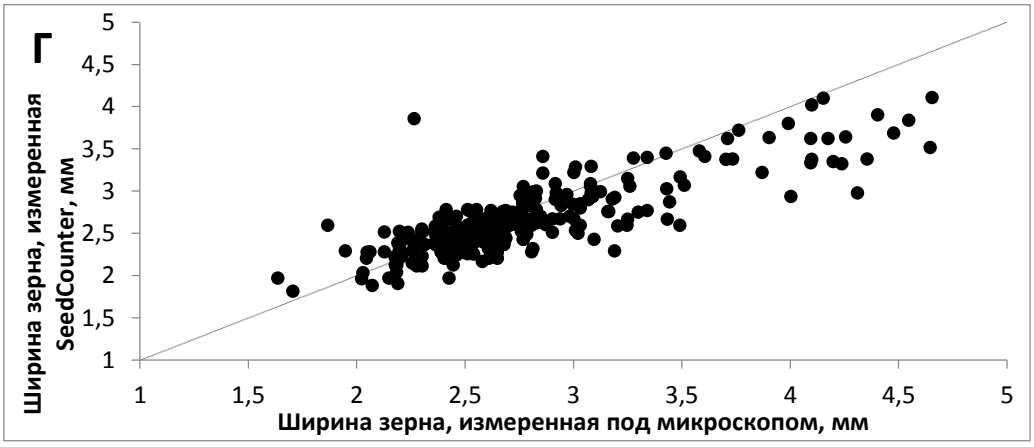

Г
Ширина зерна, измеренная SeedCounter, мм
Ширина зерна, измеренная под микроскопом, мм

Рисунок 16. Диаграмма рассеяния размеров зерен, измеренных мобильным устройством Samsung (ось Y) относительно размеров, измеренных под микроскопом (ось X). А) длина зерен при условиях L3; (Б) длина зерен в условиях L6; В) ширина зерен в условиях L3; Г) ширина зерен в условиях L6.

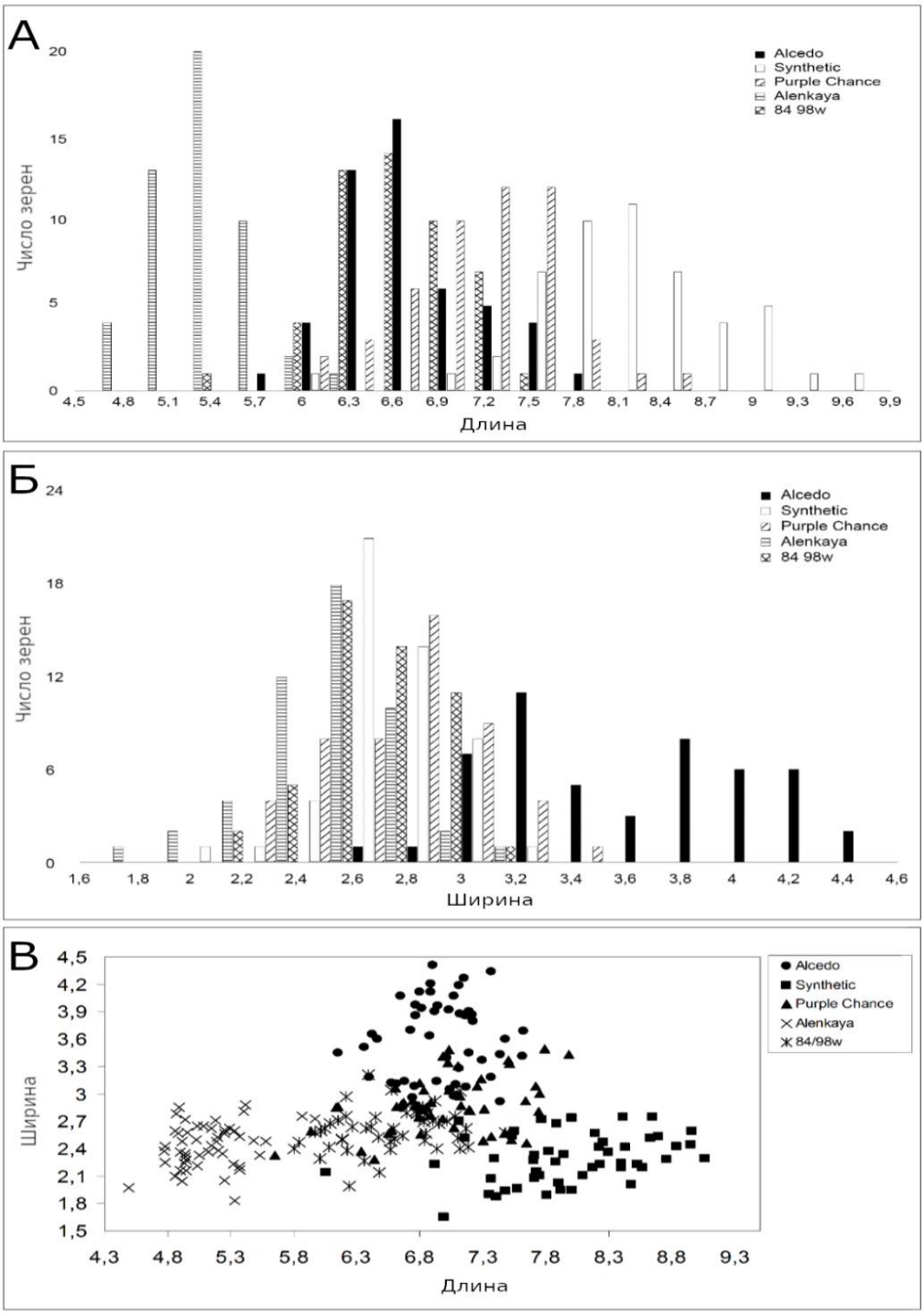


Рисунок 17. Распределение длины и ширины зерен для пяти различных сортов пшеницы. А) Гистограмма распределения длины. Б) Гистограмма распределения ширины. В) Диаграмма рассеяния зерен по длине и ширине.

3.5 Заключение по главе 3

Для оценки количественных характеристик формы и размеров зерен пшеницы на основе цифровых изображений было разработано приложение SeedCounter как для настольных ПК, так и для устройств под управлением ОС Android. Проведенные серии тестов показали, что разработанный метод подсчета морфометрических параметров зерен пшеницы, реализованный в виде мобильного и настольного приложения SeedCounter, позволяет производить сбор, анализ и передачу данных как в лабораторных, так и полевых условиях. Метод может быть использован в массовых селекционно-генетических экспериментах.

Мобильное приложение SeedCounter для мобильных устройств на базе Android можно бесплатно загрузить в Android Play Store (https://play.google.com/store/apps/details?id=org.wheatdb.seedcounter). Для приложения SeedCounter требуется минимальный уровень Android API 14 версии. SeedCounter использует библиотеку OpenCV для обработки изображений и распространяется под лицензией BSD (Berkley Software Distribution).

ГЛАВА 4. МЕТОД МОРФОМЕТРИИ КОЛОСА ПШЕНИЦЫ

С помощью анализа цветных двухмерных изображений могут быть извлечены такие важные для селекционера характеристики колоса как длина, ширина, параметры формы и др. При ручном фенотипировании, селекционеры пользуются линейкой или штангенциркулем, что позволяет делать измерения с точностью до десятых долей миллиметра. Для достижения сопоставимой точности при автоматизированной морфометрии, цифровые изображения должны быть получены в приемлемых условиях, таких как студийная съемка. Условия освещения в студиях позволяют минимизировать искажения, связанные с тенями при съемке объемных объектов на плоскости. Помимо этого, для более точного выделения контуров снимаемых объектов зачастую необходим специальная подложка-фон. Тем не менее, автоматизированная морфометрия колосьев может быть выполнена и в полевых условиях. Важным условием при таком подходе является соблюдение заранее оцененных условий освещения и точности камеры устройства.

4.1 Протоколы получения изображений

Для получения изображений, подходящих для автоматизированного анализа, были предложены два протокола съемки. Колосья фотографируются на синим фоне. Колос располагается или на предметном столе (протокол «на столе», рисунок 18 А) или закрепляется вертикально на штативе (протокол «на прищепке», рисунок 18 Б).

Рисунок 18. Изображения колосьев, полученные по двум различным протоколам. (А) На столе. Колос и цветовая шкала (ColorChecker) располагается на синем фоне на столе. Фотосъемка производится вертикально сверху-вниз. (Б) На прищепке. Колос и цветовая шкала (ColorChecker) закрепляется на штативе с помощью прищепки в вертикальном положении. Позади закреплен, также вертикально расположенный синий фон.

Штатив, используемый в протоколе «на прищепке», позволяет закреплять колос под разным углом вращения относительно своей оси. В область каждого кадра помещается цветовая шкала ColorChecker Mini Classic target (https://xritephoto.com/camera), которая позволяет произвести цветовую коррекцию изображения. Процедура цветокоррекции устраняет искажения цвета на изображении, возникающие из-за разных условий освещения (Berry et al., 2018). Другим преимуществом использования цветовой шкалы является ее стандартный размер 63 x 108 мм, что позволяет оценить масштаб изображения.

24-битные RGB изображения колосьев были получены с помощью цифровой фотокамеры в формате jpg. Примеры изображений колосьев показаны на рисунке 18.

Фотографировался основной колос каждого растения. Была предоставлена серия изображений: одно изображение с использованием протокола «на столе» и четыре с использованием протокола «на прищепке» в четырех соответствующих проекциях: две боковые и две лицевые стороны.

Общее количество полученных изображений составило 1245.

4.2 Идентификация колоса и остей на изображении

4.2.1 Предварительная обработка изображения

Обработка изображений осуществляется с использованием пакета OpenCV (Open Source Computer Vision Library, https://opencv.org) (Kaehler and Bradski, 2016). На первом этапе для снижения уровня шума на исходных изображениях применяется фильтр Гаусса. В сравнении с медианным фильтром он позволяет точнее сохранять границы колосьев.

4.2.2 Распознавание цветовой шкалы

Цветовая шкала распознается на изображении для последующей нормализации цветов (цветокоррекции), а также определения масштаба. Масштаб изображения

рассчитывается как отношение реальной площади цветовой шкалы к ее площади на изображении (с учетом поправки на ее перспективу). Алгоритм распознавания цветовой шкалы и цветокоррекции был реализован Смирновым Н.В (Genaev et al., 2019).

4.2.3 Сегментация

Сегментация изображения на области фона и колоса с остями производится путем бинаризации. Предварительно изображение преобразуется в цветовое пространство HSV, в котором значения тона, насыщенности и яркости пикселей разделены на отдельные каналы. Это соответствует естественным изменениям характеристик изображения при различном уровне освещения и параметров светочувствительности камеры и позволяет устойчивее выделять объекты на изображения на основе их натурального цвета.

Диапазон значений каналов HSV модели, соответствующих колосу, был подобран заранее (см. раздел 4.2.5). В дальнейшем значения этих границ использовались для классификации сигналов в каждом пикселе изображения. Сигналам в пикселях, где значения попадали в диапазон, присваивали значение интенсивности 1, иначе - нулевое. Цвета подбирались на основе вручную размеченных изображений из обучающей выборки, с помощью генетического алгоритма (Whitley, 1994). В случае, если пиксель изображения попадает хотя бы один из диапазонов значений набора цветов, он определяется как пиксель колоса, иначе – как пиксель фона. На полученном изображении определяются замкнутые контура функцией OpenCV findContours(…) (Suzuki et al., 1985) и выбирается самый большой по площади контур, который соответствует колосу и остям. Незначительная часть полезной информации при этом, в основном касающаяся остей, иногда теряется. Происходит это из-за малой толщины остей, которые отделяются от основной части колоса в узких местах. Меньшие по размеру контуры соответствуют мусору и фрагментам чешуек (рисунок 19 Г) и исключаются из дальнейшего анализа.

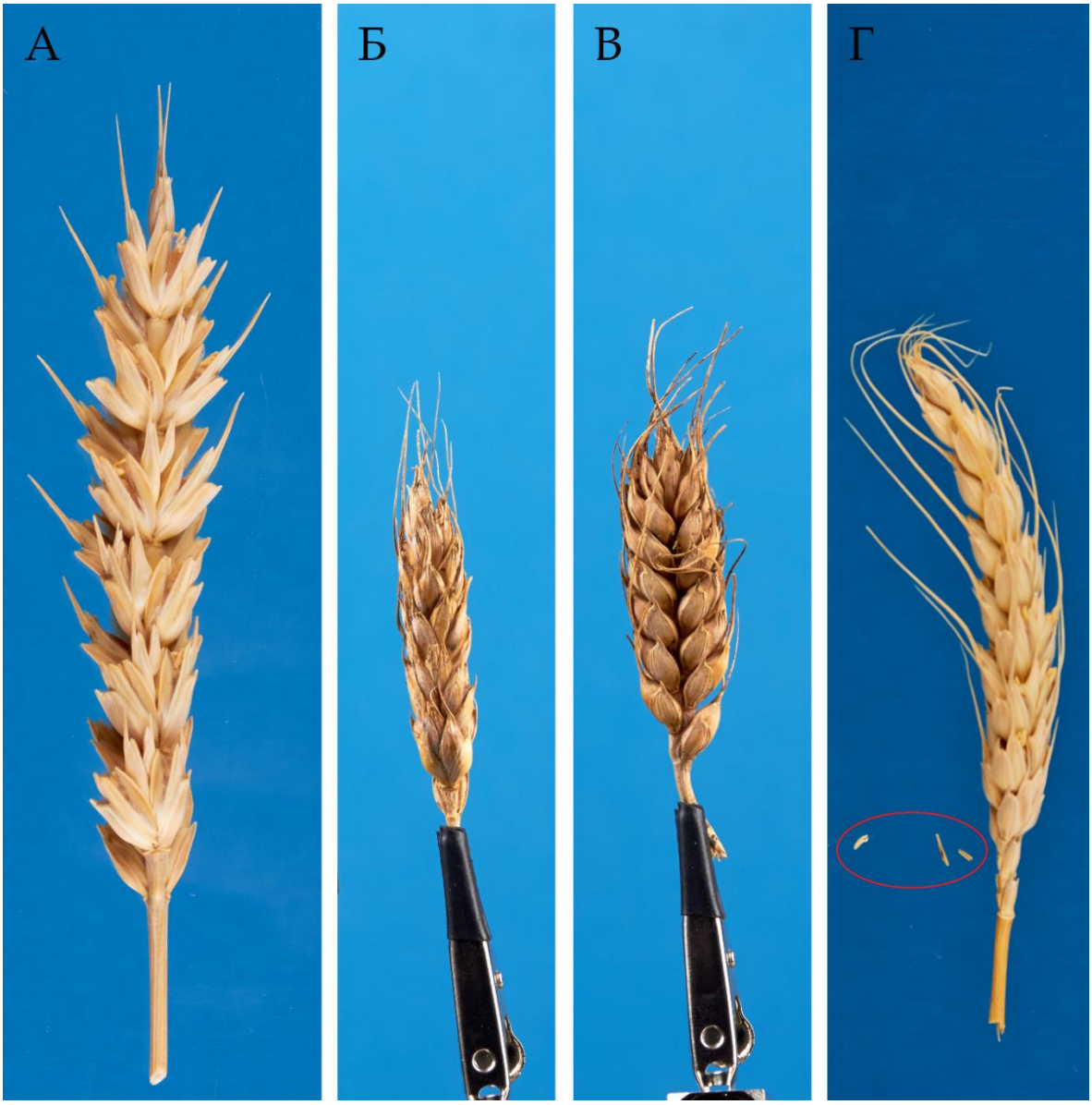


Рисунок 19. Примеры изображений колосьев, с разным типом остистости. А) безостый колос Б) колос с зачатками остей В) полу-остистый колос Г) короткоостый колос. Красным овалом отмечен мусор: фрагменты остей, чешуек.

4.2.4 Идентификация остей

Цвет остей близок к цвету колоса. На некоторых изображениях ости интенсивно пересекаются или даже слипаются в пучки. Эта особенность существенно затрудняет идентификацию отдельных остей, а в некоторых случаях делает ее невозможной. В работе предложен алгоритм сегментации остей и тела колоса, состоящий из двух этапов.

На первом этапе идентифицировались пиксели остова остей. Поскольку толщина остей существенно меньше, чем толщина тела колоса, то для получения остова остей использовался алгоритм частичной скелетизации (рисунок 20).

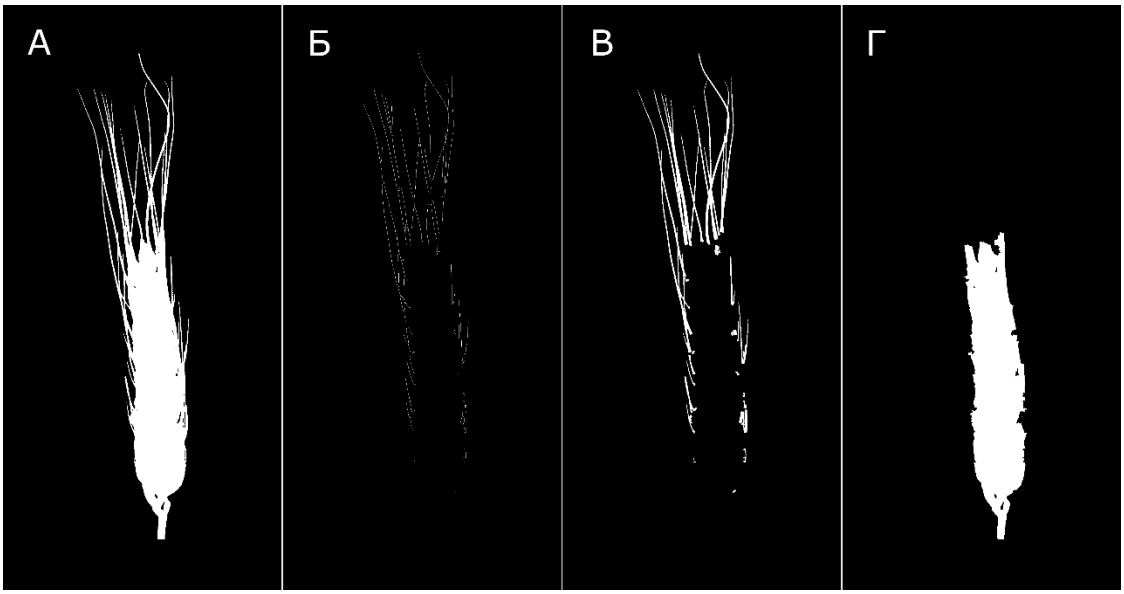


Рисунок 20. Визуализация этапов алгоритма частичной скелетизации для определения областей остей и «тела» колоса. А) Исходный отпечаток колоса с остями. Б) Скелетизированный остов остей. В) Идентифицированные области остей. Г.) Тело колоса без остей.

Данный алгоритм заключается в последовательном удалении граничных пикселей областей переднего плана, при условии сохранения их связности. Пиксели удалялись итерационно, “слоями”: сначала первый слой, потом второй, третий и т.д. Если удаление очередного пикселя слоя приводило к увеличению изолированных областей отпечатка колоса, то он отмечался как пиксель остова (не удалялся). Для остановки алгоритма использовалось пороговое ограничение на количество итераций, которое подбиралось путем оптимизации алгоритма на обучающей выборке изображений (см. раздел 4.2.5). На втором этапе определялись области остей колоса. На каждой итерации первого этапа удаляемые пиксели запоминалась. Каждый пиксель первого слоя отмечался уникальным номером. Пиксели последующих слоев отмечались номерами соседних удаленных пикселей из предыдущего слоя так, чтобы они образовывали обособленные последовательности от первого слоя, до последнего. На основе этого можно было определить пиксели,

удаление которых приводило к формированию остова - они отмечались как пиксели остей.

4.2.5 Выбор параметров для выделения областей колоса и остей на изображении

Для подбора этих параметров использовалась выборка изображений 93 колосьев гибридов F2 австралийского сорта мягкой пшеницы Triple Dirk с образцом KU506 китайской пшеницы *T. yunnanense*, полученных как по протоколу «на столе», так и «на прищепке». Для каждого колоса, на основании экспертной оценки, приведена классификация по типам остистости: безостые (рисунок 19 А), с зачатками остей (рисунок 19 Б), полу-остистые (рисунок 19 В) и короткоостые (рисунок 19 Г). Распределение изображений колосьев в этой выборке по типам протокола и остистости приведено в таблице 10. На каждом изображении вручную были маркированы пиксели остей и пиксели тела колоса. Изображения были произвольным образом разделены на тестовую (30 изображений) и обучающую (63 изображений) выборки так, что соотношение типов остистости колосьев и протоколов в тестовом и обучающем наборе было приблизительно одинаковым.

Стоит отметить, что в обучающей и тестовой выборке были колосья преимущественно обычного цветового диапазона: от темно-коричневого до светло желтого. В то время как среди различных видов пшениц встречаются колосья “экзотических” оттенков: синих, черных. В таких случаях подобранные параметры сегментации могут оказаться неоптимальными. Для анализа видов, имеющих колосья нестандартных цветов следует подобрать новые параметры сегментации на расширенной выборке (генетическим алгоритмом). Кроме того, учитывая цветовой диапазон получившейся выборки, возможно, следует заменить фон на другой, более подходящий (контрастный) цвет: белый, черный и т.п.

Таблица 10. Распределение изображений по типам остистости колосьев.

Тип остистости	**Общее Количество изображений**	**Количество изображений, полученных по Протоколу “на прищепке”**	**Количество изображений, полученных по Протоколу “на столе”**
Безостые	16	10	6
С зачатками остей	4	4	0

Полу-остистые	14	14	0
Короткоостые	59	36	23

Рассчитывались следующие индексы Жаккара:

- J_e – точность бинаризации всего колоса вместе с остями;
- J_b – точность распознавания пикселей тела колоса (область колоса за вычетом пикселей, принадлежащих остям);
- J_a – точность распознавания остей.

В процессе подбора параметров оптимизировалось средние значения индексов Жаккара для тестовой выборки с помощью генетического алгоритма (Whitley, 1994).

Блоки параметров (особи) состояли из наборов 7 цветов, соответствующих колосьям в модели HSV и диапазонов отклонений их значений (dH, dS, dV). Блоки могли обмениваться целевыми цветами со связанными диапазонами. Размер популяций от 20 до 100 особей. В результате были подобраны оптимальные целевые цвета для сегментации. Полученные оценки точности сегментации описаны в разделе 4.4.

4.2.6 Идентификация контура колоса и его выпрямление

После удаления остей в местах их сочленения с “телом” колоса иногда образовывались неправдоподобно резкие изгибы контура. Поэтому требовалось выполнить его сглаживание. С этой целью для исходного контура тела колоса вычислялись 70 элементов разложения эллиптических дескрипторов Фурье (Kuhl and Giardina, 1982), на основе которых после был восстановлен сглаженный контур.

Осевая линия колоса аппроксимировалась ломаной линией. Ломанная в свою очередь строилась на основе сегментов контура “тела” колоса. Построение этой линии выполнялось рекурсивно (рисунок 21).

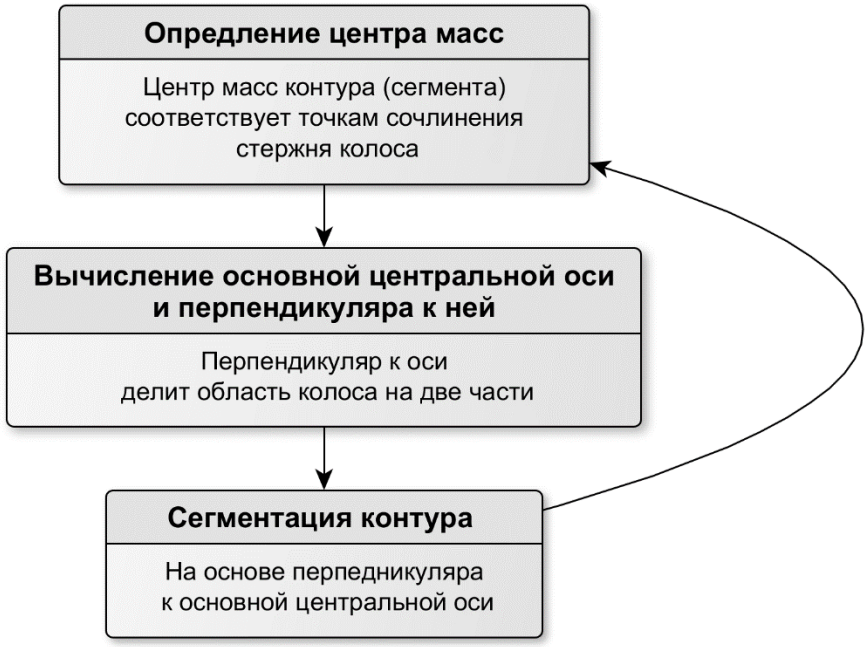


Рисунок 21. Блок-схема алгоритма построения осевой линии колоса. Алгоритм рекурсивный: начинается с определения центра масс контура, далее вычисляется основная центральная ось и перпендикуляр к ней, затем выполняется сегментация контура на основе вычисленных прямых. После, алгоритм выполняется для контуров полученных сегментов. Выход по достижению заданного числа сегментов.

На первом шаге для пикселей контура определялся центр масс (среднее геометрическое координат точек контура) и главные оси эллипсоида, аппроксимирующего контур. Основная центральная ось соответствовала положению стержня колоса в центре масс. Перпендикуляр к большей основной центральной оси от центра масс контура делит область колоса на две части. На втором шаге аналогичная процедура выполнялась для каждой из этих частей. В результате, каждая из них также делилась соответствующей осью на две части, для которых выполнялся следующий шаг итерации. Данная процедура выполнялась последовательно, пока количество сегментов не превышало 20. Центры масс каждого из полученных сегментов определяли ломанную, аппроксимирующую осевую линию колоса. На последних этапах итерации в некоторых случаях размеры сегмента поперек оси могли быть больше, чем вдоль нее. Поэтому основная центральная ось такого сегмента оказывалась перпендикулярной осевой линии колоса. В этом случае из двух осей сегмента выбиралась та, которая была

сонаправлена основной центральной оси сегмента, полученного на предыдущем этапе.

После определения осевой линии производится процедура выпрямления контура колоса. Для этого, геометрически определялись “ребра тела” колоса: отрезки, соединяющие осевую линию и точки контура колоса. Отрезки строились слева и справа от осевой линии, перпендикулярно ей, формируя соответствующие полупрофили, начиная с нижней точки осевой линии и двигаясь к вершине с шагом в 1 пиксель.

Ребра тела колоса формируют его профиль – независимый от изогнутости колоса образ. Отрисованный на чистом изображении профиль колоса позволяет получить выпрямленный контур при условии, что отрезки осевой линии раскладываются на одной прямой. Расстояния пикселей трансформированного контура до центральной линии равны расстояниям соответствующих пикселей на исходном контуре. Такое преобразование позволяет устранить деформации контура колоса, вызванные его изгибом. Определение размера колоса и его количественные характеристики формы вычисляются на изображении выпрямленного колоса.

4.2.7 Интегральные характеристики формы

К интегральным характеристикам формы колоса относятся: длина колоса (L_e, измеряется в миллиметрах), которая аппроксимируется длиной осевой линии колоса; периметр контура колоса без остей (P_e, измеряется в миллиметрах); площадь колоса (S_e, измеряется в миллиметрах квадратных); отношение площади колоса к квадрату его длины (SQI); округлость (C); индекс закругленности (R); индекс шероховатости (Rg); индекс компактности/целостности (S).

$$SQI = \frac{S_e}{L^2} \tag{6}$$

Округлость отражает насколько форма контура близка к форме окружности. Она вычисляется по формуле 7. Значение округлости варьируется от 0 до 1. При этом значение 1 означает, что объект имеет форму идеального круга.

$$C = \frac{4\pi \times area}{perimeter^2} \quad (7)$$

, где area – площадь контура, perimeter – периметр контура.

На контурах, имеющих много небольших выпуклостей на поверхности, наблюдается увеличение значения периметра и уменьшение значения округлости. В этих случаях целесообразно использовать индекс закругленности (roundness, R). Так как он инвариантен к неровностям контура:

$$R = \frac{4 \times area}{\pi\,[Major\ axis]^2}, \quad (8)$$

где area – площадь контура, Major axis – длина главной оси контура.

Индекс шероховатости (rugosity, Rg) определяется как отношение периметра контура к выпуклому периметру:

$$Rg = \frac{Ps}{Pc}, \quad (9)$$

где Ps - периметр контура, Pc - выпуклый периметр контура. Последний параметр также известен в литературе как выпуклая оболочка и определяется как граница наименьшей выпуклой фигуры, которая содержит все точки изображения.

Индекс компактности/целостности (solidity, S) - это отношение площади контура к площади его выпуклой оболочки:

$$S = \frac{Contour\ Area}{Convex\ Hull\ Area}, \quad (10)$$

где *Contour Area* – площадь контура, *Convex Hull Area* – площадь выпуклой оболочки контура.

4.3 Модель четырехугольников

Контур колоса можно представить в виде двух четырехугольников (рисунок 22). В таком представлении осевая линия колоса является общим основанием четырехугольников.

Ребра контура, полученные на этапе выпрямления “тела” колоса, по отдельности, левые и правые, аппроксимируются двумя четырехугольниками с одной смежной стороной-основанием (рисунок 22). Остальные стороны определяется четырьмя независимыми параметрами:

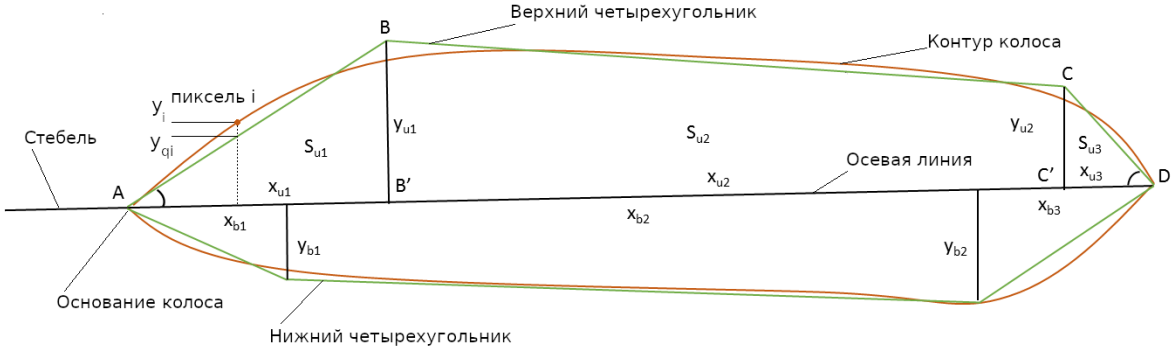


Рисунок 22. Представление формы колоса в виде двух четырехугольников. Черной горизонтальной линией показана осевая линия колоса; Контур колоса показан коричневой линией. Четырехугольники, аппроксимирующие контур колоса показаны зелеными линиями. Основные параметры, характеризующие геометрию показаны для верхнего четырехугольника. Пунктирная линия демонстрирует перпендикуляр, который строится от каждого i-го пикселя контура к осевой линии; показано значение высоты для каждого i-го пикселя y_i ребра четырехугольника y_{qi}.

x_{u1} – расстояние от вершины колоса до проекции B’ вершины B на основание AD;

x_{u2} – расстояние от B’ до проекции C’ вершины C на основание AD;

y_{u1} – расстояние от вершины B до ее проекции B’ на основание AD;

y_{u2} – расстояние вершины C до ее проекции C’ на основание AD;

Расстояние x_{u3} от проекции C’ до основания колоса D можно вычислить, зная длину колоса $x_{u3}=L_e - x_{u2}$.

Для нижнего четырехугольника определяются аналогичные параметры x_{b1}, x_{b2}, x_{b3}, y_{b1}, y_{b2} (рисунок 22).

Параметры верхнего и нижнего четырехугольника подбираются таким образом, чтобы минимизировать квадратичное отклонение по высоте между пикселями контура и ребрами четырехугольника, которое при фиксированной длине колоса зависит от четырех параметров: $D_{sq} = D_{sq}(x_{u1}, x_{u2}, y_{u1}, y_{u2})$ (для верхнего четырехугольника):

$$D_{sq} = \sum_i (y_i - y_{qi})^2$$

, где y_i – расстояние от i-й точки контура колоса перпендикулярно к осевой линии, y_{qi} – росстояние от i-й точки четырехугольника перпендикулярно к осевой линии. Минимизация производится с помощью алгоритма Левенберга-Марквардта - метода оптимизации, направленного на решение задач о наименьших квадратах (Press et al., 1992). Алгоритм реализован в библиотеке Apache (org.apache.commons.math3.fitting.leastsquares.LevenbergMarquardtOptimizer, http://commons.apache.org/proper/commons-math/javadocs/api-3.4/org/apache/commons/math3/optimization/general/LevenbergMarquardtOptimizer.html).

Помимо основных параметров модели, рассчитывается также ряд производных параметров для двух четырехугольников:

- S_{u1} – площадь треугольника ABB’ (mm2);
- S_{u2} – площадь трапеции BB’C’C (mm2);
- S_{u3} - площадь треугольника ABB’ (mm2);
- S_u– площадь верхнего четырехугольника (mm2);
- y_{um} – среднее значение высоты верхнего четырехугольника (mm);

Аналогично вычислялись параметры для нижнего четырехугольника.

Помимо этого, для двух четырехугольников вычислялись:

$AI_{x2}=(x_{u1}-x_{b1})2/x_{u1}+(x_{u2}-x_{b2})2/x_{u2}+(x_{u3}-x_{b2})2/x_{u3}$ – индекс асимметрии по длинам сегментов (mm);

$AI_{y2}=(y_{u1}-y_{b1})2/y_{u1}+(y_{u2}-y_{b2})2/y_{u2}$ – индекс асимметрии по высотам сегментов (mm)

$AI_{xy2}= AI_{x2}+AI_{y2}$ – общий индекс асимметрии (mm).

4.4 Оценка точности распознавания областей остей и колоса

Для оценки работоспособности предложенного подхода морфометрии колосьев, из созданной базы данных SpikeDroidDB были взяты изображения колосьев для 249 растений, которые были проаннотированы вручную. Использовались оцифрованные данные 1245 фотографий колосьев, полученных для 9 различных видов и гибридов гексаплоидных пшениц (таблица 3). Эти виды пшениц имеют контрастную форму колоса, которая контролируется хорошо изученными генами (Swaminathan and Rao, 1961). Для анализа у каждого растения фотографировался главный колос. Общее количество фотографий для отдельного колоса составляет 5 штук. Они получены при разных протоколах: в одной проекции по протоколу «на столе» и четырех проекциях по протоколу «на прищепке».

Среднее значение точности методов распознавания тела колоса и остей составило J_b = 0,925 и J_a = 0,660, соответственно, после подбора оптимальных параметров алгоритма сегментации на тестовой выборке изображений. На обучающей выборке среднее значение точности методов оценивается как J_b = 0,932 и J_a = 0,634.

Использование цветокоррекции не оказало существенного влияния на точность распознавания объектов. Среднее значение индекса Жаккара J_b для сегментации “тела” колоса с учетом цветокоррекции составило 0,925, а для остей J_a – 0,679.

На рисунке 23 приведен в качестве примера изображение колоса с идентификатором “6450” из БД SpikeDroidDB и результаты отдельных этапов цифровой обработки. Для данного изображения была получена оценка точности распознавания колоса J_b=0,963 и остей J_a=0,796. На изображении с выделенными остями заметно, что основная доля неразмеченных пикселей остей концентрируется на их кончиках, а для большей части остей их пиксели правильно идентифицировались.

Рисунок 23. Определение областей остей на изображении колоса вида T. yunnanense. (А) Исходное изображение колоса на синем фоне. (Б) Бинаризованный отпечаток колоса. (В) Исходное изображение колоса с отмеченными областями остей (красные линии) и ребрами профиля (синие линии).

Была проведена оценка влияния различных факторов, связанных с протоколом получения изображения и его обработки, на точность определения остей. В качестве таких факторов выделили масштаб съемки (количество пикселей на единицу фотографируемой площади), тип протокола («на столе» или «на прищепке»),

проекция колоса (лицевая или боковая сторона колоса) для протокола «на прищепке».

Прежде всего, было исследовано влияние на точность идентификации остей масштаба изображения. В использованной выборке изображений, для оценки точности выделения контуров оказались 4 масштаба (рисунок 24), которые определялись расстоянием от объектива камеры до плоскости расположения колоса. При этом изображения, полученные по протоколу «на столе» расположены вдоль одной линии (масштаб 4, оранжевые кружки).

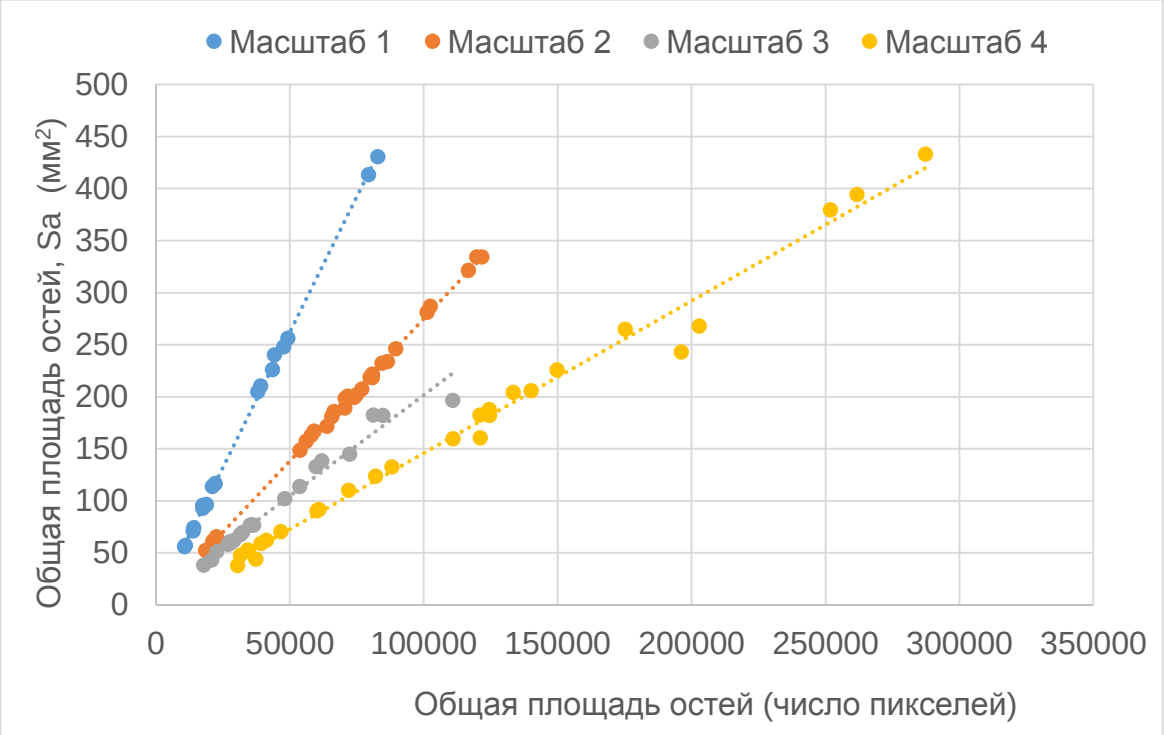


Рисунок 24. Отношение реальной площади остей колосьев к числу пикселей областей остей. По вертикальной оси отмечена общая площадь остей в мм^2. По горизонтальной оси – общая площадь остей в пикселях. Цветом и типом меток обозначены различные масштабы съемки.

Остальные величины масштаба соответствуют изображениям, полученным по протоколу «на прищепке». Таким образом, при применении предложенных протоколов различия в масштабе изображений являются неизбежными и использование цветовой шкалы для определения масштаба изображения является оправданным.

Распределение параметра индекса Жаккара J для идентификации “тела” колоса и остей для всех 93 изображений приведено на рисунке 25.

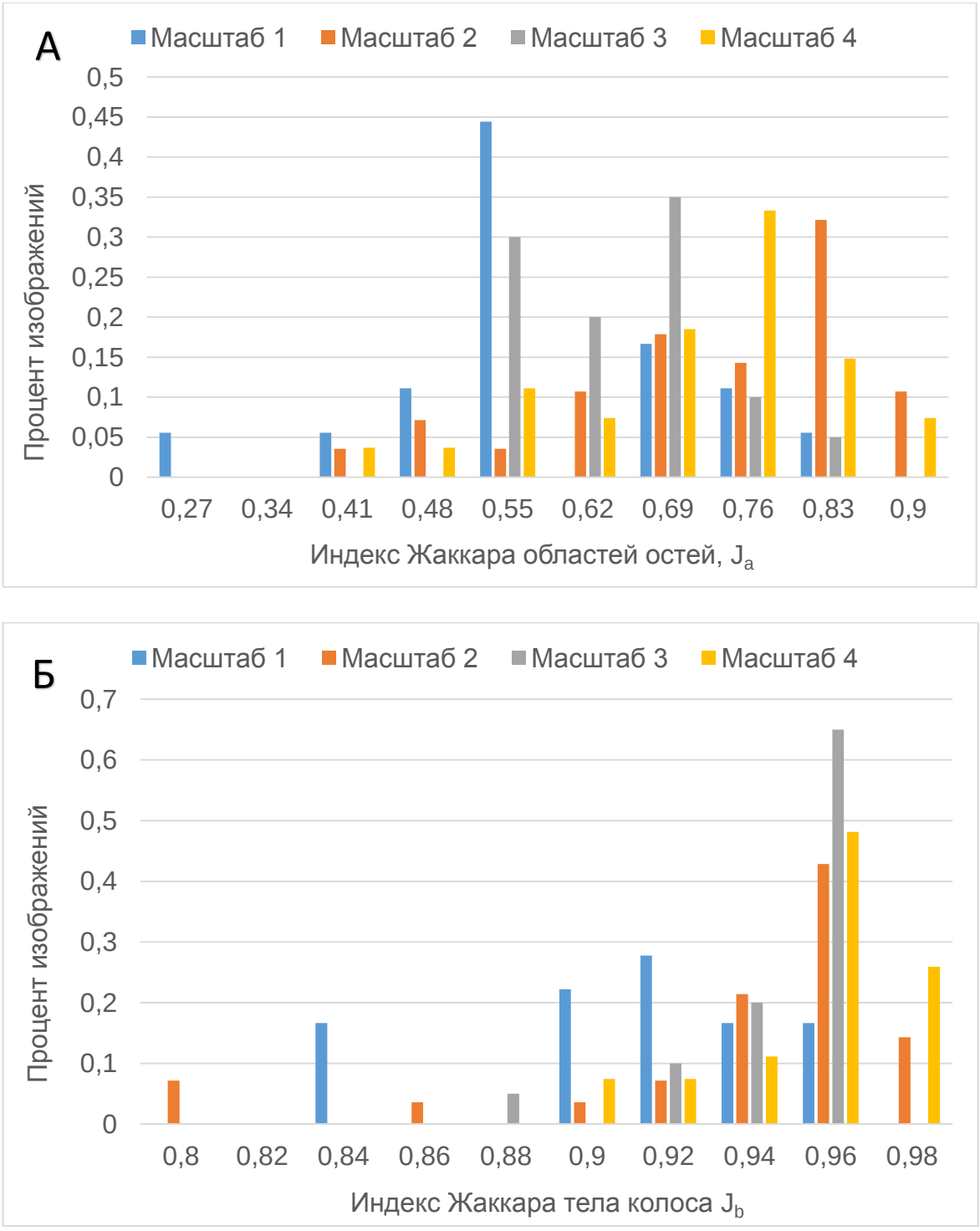


Рисунок 25. Распределение индекса Жаккара для остей (А) и "тела” (Б) колосьев, сгруппированных по масштабам съемки. На горизонтальной оси отмечены значения индекса Жаккара (среднее значение бина). По вертикали отмечен процент изображений колосьев, с данным значением индекса Жаккара (попавших в данный бин).

Оценка точности распознавания областей остей J_a находится в интервале от 0,27 до 0,9. Для масштаба 1, распределение точности смещено в меньшую сторону (рисунок 25) и в среднем составило около 0,549. Это объясняется тем, что масштаб 1 самый крупный масштаб съемки в эксперименте (наименьшее расстояние от объектива до объекта). Для остальных масштабов 2, 3 и 4 распределения J_a слабо различаются между собой и средние значения J_a равны 0,695, 0,607 и 0,676, соответственно. Оценка точности распознавания «тела» колоса J_b распределилась в интервале от 0,8 до 0,98. Средние значения составили: 0,901, 0,928, 0,938, 0,945 для 1, 2, 3, 4 масштабов, соответственно. Можно заметить, что для параметра J_b наблюдается явная тенденция увеличения точности распознавания области «тела» колоса при уменьшении масштаба, в отличие от параметра J_a. При уменьшении размера пикселя повышается детализация контура тела колоса, но для областей этого не происходит, вероятно, из-за соотношения периметра и площади.

Вторым возможным фактором, влияющим на точность работы алгоритма, является тип протокола. Для протокола «на столе», в серии снимков расстояние от камеры до объекта фиксируется. Колос расположен горизонтально на поверхности стола и находится в плоскости, перпендикулярной оси объектива. Для протокола «на прищепке» колос может отклоняться от плоскости, перпендикулярной оси объектива в силу его изогнутости или случайных отклонений оси от вертикали при установке колоса на прищепке. Было проанализировано распределение индекса Жаккара J для изображений, сгруппированных по протоколам (рисунок 26).

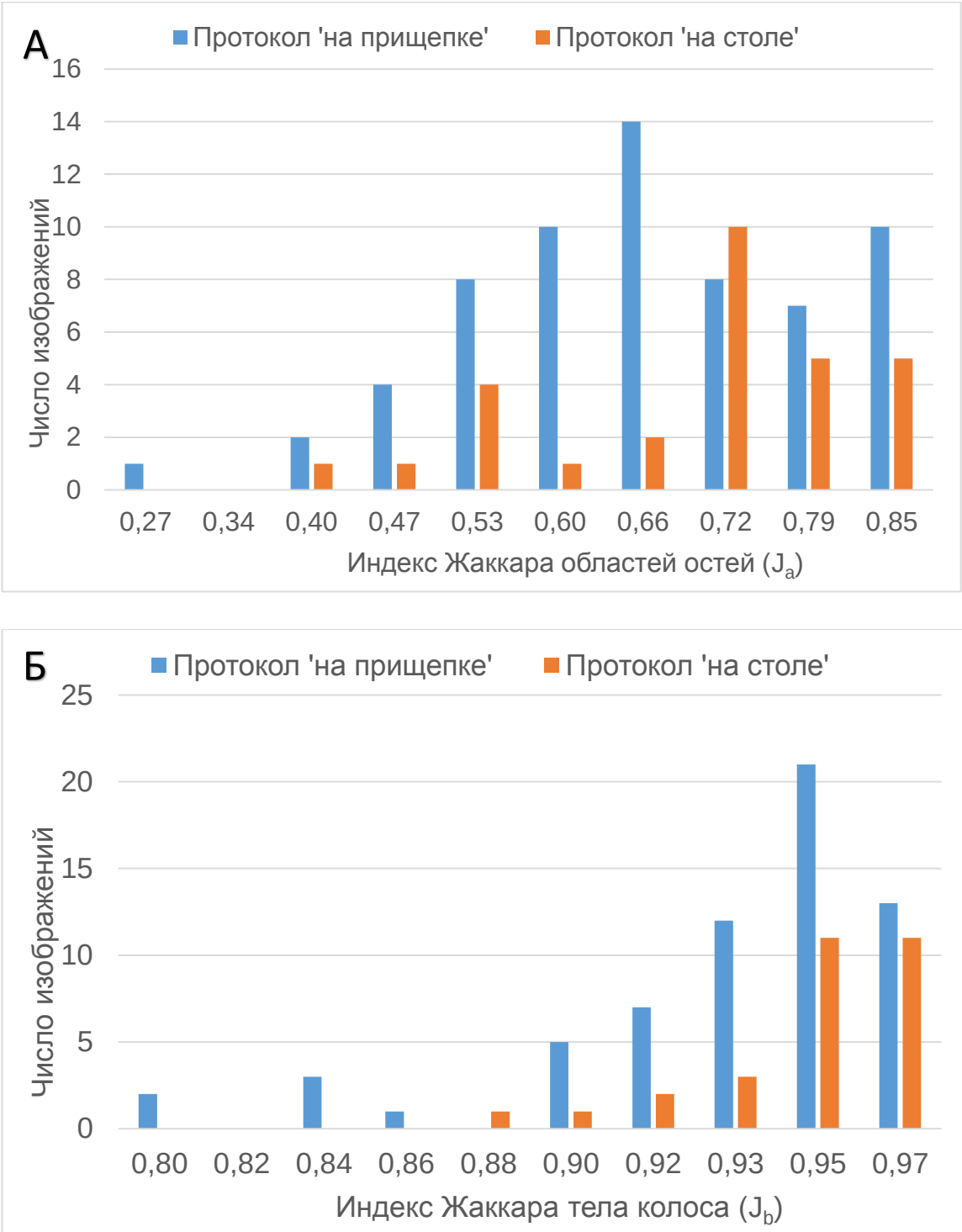


Рисунок 26. Распределение значений индексов Жаккара для остей колосьев (А) и тела колоса (Б) сгруппированные по протоколам «на столе» и «на прищепке». Чем больше индекс Жаккара, тем ближе автоматическая сегментация к разметке, выполненной вручную.

График демонстрирует, что масштаб изображения значимо влияет на точность распознавания как тела колоса, так и остей. Тип протокола оказывают существенное влияние на точность определения тела колоса, но не остей.

4.4.1 Анализ параметров остистости для выборки колосьев

Были проанализированы значения площади остей (в мм^2), S_a. измеренные программой. Значимые различия распределений S_a на выборке из 46-ти изображений (без повторов в разных проекциях) были выявлены при сравнении изображений колосьев с вариантами остистости “с зачатками остей” и “короткоостые” (χ^2= 22,64; p<0,05). Также значимо различались распределения у колосьев с типами остистости “безостые” и “короткоостые” (χ^2=24,75; p<0,05); “короткоостые” и “полу-остистые” (χ^2= 18,09; p<0,05). Из этого следует, что оценки на основе анализа изображений соответствуют экспертным оценкам, что в свою очередь подтверждают работоспособность метода.

Распределение представлено на рисунке 27. Из рисунка видно, что самую большую дисперсию площади остей имеют колосья с типом остистости “короткоостые”, которые имеют как малые, так и большие площади остей. Пик количества изображений колосьев с данным типом остистости приходится на средние площади ($180 < S_a < 222$ мм^2). Наименьшую площадь остей имеют колосья с типом остистости “с зачатками остей”. Далее по возрастанию площади остей идут “безостые” колосья и “полу-остистые” колосья. Однако, полученные распределения значений S_a для разных типов колосьев достаточно сильно пересекаются. Можно выделить только один интервал площади S_a (S_a >264 мм^2), который содержит колосья только одного типа - короткоостых. Таким образом, одна лишь площадь не может использоваться для точной идентификации типов колосьев.

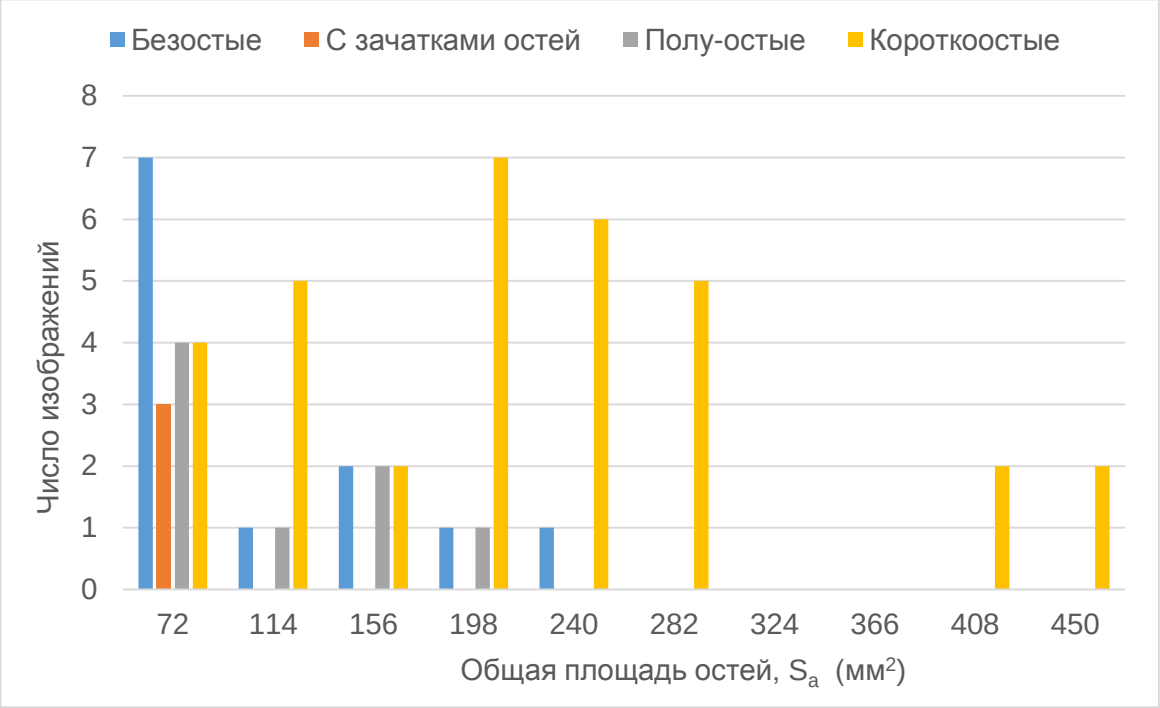


Рисунок 27. Гистограмма распределения площади остей разных типов колосьев. По горизонтальной оси отложено значение общей площади остей в мм^2 (значение бина). По вертикальной оси отложено число изображений колосьев с данным значением площади остей (попавших в данный бин).

4.5 Анализ характеристик формы колоса

С помощью разработанного нами метода фенотипирования мы оценили количественные характеристики формы и размера колосьев для четырнадцати генотипов пяти видов гексаплоидных пшениц (всего 160 образцов). Использовались изображения, полученные по протоколу «на столе».

Целью анализа было оценить разнообразие количественных характеристик колоса для исследованных образцов растений и определить вариабельность количественных характеристик колоса для растений одного генотипа и между генотипами. Помимо этого, необходимо было определить зависимость между характеристиками колосьев и на основании этого выделить их группы. Это позволило оценить адекватность предложенного метода при сравнении полученных результатов с известной систематикой пшениц.

Для приведения параметров формы колоса в симметричный вид, часть парных параметров модели четырехугольников были объединены в суммирующие, либо

усредняющие параметры. Так параметры определяющие координаты проекции на осевую линию колоса были объединены в усредняющие параметры. Например, quadrangle_x1 для объединения моделей вычислялся как (quadrangle_x1 + quadrangle_x1') / 2, т.е. среднее арифметическое от параметров quadrangle_x1 первого четырехугольника и quadrangle_x1' второго четырехугольника. Аналогично рассчитывались объединенные параметры: quadrangle_x2, quadrangle_x3, а также их нормированные на длину колоса значения: quadrangle_x1_norm, quadrangle_x2_norm.

Параметры, определяющие ширину колоса, были объединены в суммирующие параметры. Так, например, параметр quadrangle_y1 для объединения моделей вычислялся как сумма параметров quadrangle_y1 первого четырехугольника и quadrangle_y1' второго четырехугольника. Аналогично были вычислены остальные суммирующие параметры: quadrangle_y2, quadrangle_S1, quadrangle_S2, quadrangle_S3, quadrangle_S, quadrangle_ym, quadrangle_h1_norm и quadrangle_h2_norm. Суммарная информация о полученных параметрах: краткие обозначения, названия и тип, приведены в таблице 11.

Таблица 11. Краткие обозначения полученных и усредненных характеристик колосьев. Приведено обозначение характеристики, его название и категория: общий параметр формы, параметр модели четырехугольников или производный параметр.

Краткое обозначение	**Название**	**Общий/Модель/ Производный параметр**
q_x1	Высота сегмента основания колоса	Модель четырехугольников
q_x2	Высота центрального сегмента колоса	Модель четырехугольников
q_x3	Высота вершинного сегмента колоса	Модель четырехугольников
q_x1_n	Нормированная высота сегмента основания колоса	Модель четырехугольников
q_x2_n	Нормированная высота центрального сегмента колоса	Модель четырехугольников
q_x3_n	Нормированная высота вершинного сегмента колоса	Модель четырехугольников
q_y1	Ширина колоса у основания	Модель четырехугольников
q_y2	Ширина колоса у вершины	Модель четырехугольников
q_y1_n	Нормированная ширина колоса у основания	Модель четырехугольников
q_y2_n	Нормированная ширина колоса у вершины	Модель четырехугольников

q_S1	Площадь сегмента основания колоса	Модель четырехугольников
q_S2	Площадь центрального сегмента колоса	Модель четырехугольников
q_S3	Площадь вершинного сегмента колоса	Модель четырехугольников
S1S	Нормированная площадь сегмента основания колоса	Нормированный параметр = q_S1/q_S
S2S	Нормированная площадь центрального сегмента колоса	Нормированный параметр = q_S2/q_S
S3S	Нормированная площадь вершинного сегмента колоса	Нормированный параметр = q_S3/q_S
q_S	Площадь модели четырехугольников	Модель четырехугольников
q_ym	Нормированная площадь	Производный = q_S/q_L
q_L	Длина колоса	Модель четырехугольников
c_Perimeter	Периметр колоса	Общий
c_Ear area	Площадь колоса	Общий
c_Awns area	Площадь остей колоса	Общий
c_Circularity	Индекс округлости контура колоса	Общий
c_Roundness	Индекс закругленности контура колоса	Общий
c_Solidity	Индекс целостности контура колоса	Общий
c_Rugosity	Индекс шероховатости контура колоса	Общий

Мы оценили среднее и стандартное отклонение некоторых основных характеристик для каждого из генотипов. Полученные значения приведены в таблице 12. Наибольшую длину колоса L, порядка 10-12 см, имеют генотипы Rother Sommer Kolben/к-1731 и к-19092 вида T. *spelta*. Немного короче колосья генотипов Новосибиская 67/NSK67 и Бабило вида T. *aestivum* – 8-10 см. Генотипы же WAG 8326 (T. *compactum*), к-14976 и к-33750 (T. *sphaerococcum*), к-56397 (T. *antiquorum*) имеют наименьшую длину – 4,5 – 5,2 см. При средней дисперсии около 1,7 см по всем генотипам.

Наибольшую ширину у основания (параметр q_y1) и вершины (параметр q_y2) имеют колосья растений вида T. *compactum* (кроме генотипа #29/к1709 по ширине основания), а также генотипы Новосибиская 67/NSK67 вида T. *aestivum* (по ширине основания) и генотип к-14976 вида T. *sphaerococcum* (по ширине вершины).

Наименьшая ширина у основания и у вершины наблюдается у генотипа Rother Sommer Kolben/к-1731 вида T. spelta, а также у генотипов к-19092.

Генотипы Rother Sommer Kolben/к-1731 и к-19092 вида T. *spelta*, а также генотип к-33750 вида T. *sphaerococcum* имели наименьшую ширину у основания, от 4 до 6 мм. У вершины же наименьшую ширину имели генотип Новосибиская 67/NSK67 вида T. *aestivum* и генотип Rother Sommer Kolben/к-1731 вида T. *spelta*, 4,6 – 4,8 мм. Следует отметить невысокую среднюю дисперсию данных характеристик по всем генотипам: 3,1 мм для ширины у основания и 3,6 для ширины у вершины.

По площади тела колоса (параметр q_S) растения распределились следующим образом: наибольшую площадь имели генотипы к-19092 вида T. *spelta*, Бабило и Новосибиская 67/NSK67 вида T. *aestivum*, а также генотипы #29/к1709, #31/к1711 и #33/к1713 вида T. *compactum*; среднюю площадь имели генотипы #37/k-1386, АНК-23 вида T. *aestivum*, Rother Sommer Kolben/к-1731 вида T. *spelta*, WAG 8326 вида T. *compactum*, к-14976 вида T. *sphaerococcum* и к-56398 вида T. *antiquorum*.

Схожим образом ведут себя индексы формы контура колоса c_Circularity, c_Roundness, c_Solidity, c_Rugosity, а также характеристика ширины колоса у основания q_y1: низкие значения для генотипов вида T. *spelta*, высокие для генотипа WAG 8326 вида T. *compactum*, средние для #29/к1709 вида T. *compactum*. Для характеристик c_Roundness и q_y1 также отмечены высокие значения для генотипов #31/к1711 и #33/к1713 вида T. *compactum* и генотипа к-14976 вида T. *sphaerococcum*.

Для генотипов к-14976 и к-33750 вида T. *sphaerococcum*, а также к-56397 вида T. *antiquorum* выявлены низкие значения параметров q_S1, q_L, c_Perimeter и c_Awns area.

Наиболее вариабельными признаками являются характеристики площадей (q_S1, q_S2, q_S3, q_S, c_Ear area, c_Awns area), а также периметр контура колоса (c_Perimeter), обуславливающаяся размерностью данных величин. Среди прочих характеристик, наиболее вариабельными являются общая длина колоса, а также длина средней части колоса.

Таблица 12. Средние значения и стандартное отклонение основных параметров колоса, характеризующих его размер и форму.

			Среднее значение/стандартное отклонение																
Вид	Сорт/Каталоговый номер	Число образцов	q_x1	q_x2	q_x3	q_y1	q_y2	q_S1	q_S2	q_S3	q_S	q_L	c_Perimeter	c_Eararea	c_Awnsarea	c_Circularity	c_Roundness	c_Solidity	c_Rugosity
T. aestivum			28,363	45,276	10,848	7,591	7,572	122,210	326,783	48,169	497,162	84,487	242,352	547,411	28,397	0,165	0,098	0,727	1,241
	#37/k-1386	18	13,049	23,730	13,924	3,137	2,941	77,317	162,002	99,955	128,879	17,125	32,056	98,931	13,053	0,060	0,029	0,104	0,280
			36,039	34,194	13,059	6,416	9,167	125,857	230,899	71,722	428,477	83,292	244,288	462,810	13,867	0,118	0,097	0,633	1,322
	АНК-23	17	17,886	16,843	8,735	3,630	3,299	99,404	111,791	62,554	134,017	19,593	37,160	124,512	5,113	0,038	0,036	0,079	0,255
			31,856	56,098	10,034	8,304	7,275	134,162	433,834	45,347	613,343	97,989	265,021	665,367	18,514	0,146	0,094	0,763	1,202
	Бабило	10	7,510	9,866	6,161	2,268	1,366	45,635	97,786	34,319	135,418	7,798	30,652	130,002	4,124	0,029	0,014	0,056	0,087
			20,716	61,956	3,281	9,479	4,667	100,927	438,810	8,216	547,952	85,953	217,640	573,260	35,689	0,184	0,110	0,835	1,142
	Новосибиская 67/NSK67	4	3,024	19,740	2,114	1,344	1,426	23,268	151,389	5,300	174,970	23,704	48,081	168,474	20,119	0,036	0,034	0,016	0,037
T. spelta			21,937	70,155	6,962	4,335	4,825	62,170	343,852	19,375	425,396	99,054	256,314	463,042	6,990	0,103	0,057	0,696	1,159
	Rother Sommer Kolben/к-1731	7	15,135	18,836	5,009	1,764	2,840	50,634	231,789	17,353	213,951	13,475	24,073	206,512	2,489	0,041	0,012	0,117	0,099
			38,805	70,093	14,281	5,916	10,241	128,563	445,815	76,666	651,045	123,179	282,864	641,353	78,507	0,111	0,052	0,697	1,159
	к-19092	9	40,228	29,831	9,889	2,048	7,577	153,321	294,913	72,081	236,909	17,840	42,938	239,035	48,924	0,058	0,006	0,176	0,128
T. compactum			15,436	26,602	4,654	14,913	11,073	113,492	348,109	27,238	488,839	46,692	252,602	527,995	220,856	0,331	0,337	0,801	1,386
	WAG 8326	8	8,617	8,347	3,839	3,326	4,269	68,761	120,889	22,142	131,518	12,961	40,345	121,281	45,285	0,172	0,121	0,118	0,257
			33,118	36,118	6,030	6,834	13,889	150,408	365,973	46,259	562,640	75,267	254,703	634,508	109,334	0,202	0,139	0,730	1,244
	#29/к1709	19	18,379	12,879	3,141	3,374	3,082	127,300	141,909	25,262	144,364	18,529	26,552	90,105	39,007	0,052	0,062	0,047	0,218
	#31/к1711	17	34,853	29,031	8,230	9,932	12,892	186,431	324,352	68,291	579,073	72,113	225,690	590,235	74,726	0,214	0,155	0,729	1,235

			17,413	10,593	6,267	2,897	4,775	99,197	136,858	80,258	167,475	16,644	42,218	151,270	50,551	0,054	0,055	0,079	0,093
			26,580	36,900	7,421	9,080	12,606	148,915	383,518	51,689	584,122	70,901	217,290	602,507	41,535	0,234	0,155	0,772	1,198
	#33/к1713	18	15,312	11,304	4,569	3,402	2,558	118,629	135,565	37,499	117,198	11,500	30,519	127,167	21,285	0,058	0,052	0,089	0,071
T. sphaerococcum			19,031	22,649	7,723	8,920	10,247	63,530	239,196	41,769	344,496	49,403	197,901	440,453	9,093	0,244	0,229	0,716	1,321
	к-14976	9	10,606	30,099	7,585	4,847	6,058	30,199	415,575	56,487	463,585	37,102	50,116	457,137	3,106	0,152	0,143	0,162	0,283
			27,293	14,593	10,968	5,561	8,067	64,366	85,143	53,200	202,708	52,854	170,048	247,576	13,434	0,166	0,131	0,609	1,245
	к-33750	10	13,244	8,469	8,928	4,656	3,292	41,655	47,434	47,572	47,265	15,870	31,050	55,604	4,586	0,071	0,069	0,106	0,097
T. antiquorum			23,700	19,552	6,359	7,194	7,013	90,881	130,169	31,291	252,341	49,611	153,733	272,733	7,185	0,211	0,162	0,704	1,232
	к-56397	10	15,237	7,708	7,379	3,048	2,732	66,625	51,935	38,158	62,248	12,852	24,316	36,903	2,274	0,097	0,077	0,102	0,076
			39,611	23,665	14,769	7,560	9,296	144,582	196,921	99,294	440,797	78,046	217,153	502,249	10,287	0,171	0,122	0,682	1,197
	к-56398	10	20,072	12,583	13,936	3,770	4,577	117,772	105,233	107,602	65,528	18,450	25,737	32,103	2,627	0,066	0,046	0,108	0,080

4.5.1 Анализ корреляций между характеристиками

Чтобы выяснить какие группы образуют полученные параметры и какие признаки они характеризуют был выполнен кластерный анализ. Мы рассчитали коэффициенты корреляции Пирсона r между основными характеристиками колосьев в выборке по всем генотипам и с помощью меры близости $d_r=(1-r)$ провели иерархический кластерный анализ этих характеристик. Кластеризация и построение дендрограмм выполнялись в программе PAST, описанной в разделе 2.4.

Анализ выявил несколько групп, характеризующих признаки формы колоса (рисунок 28).

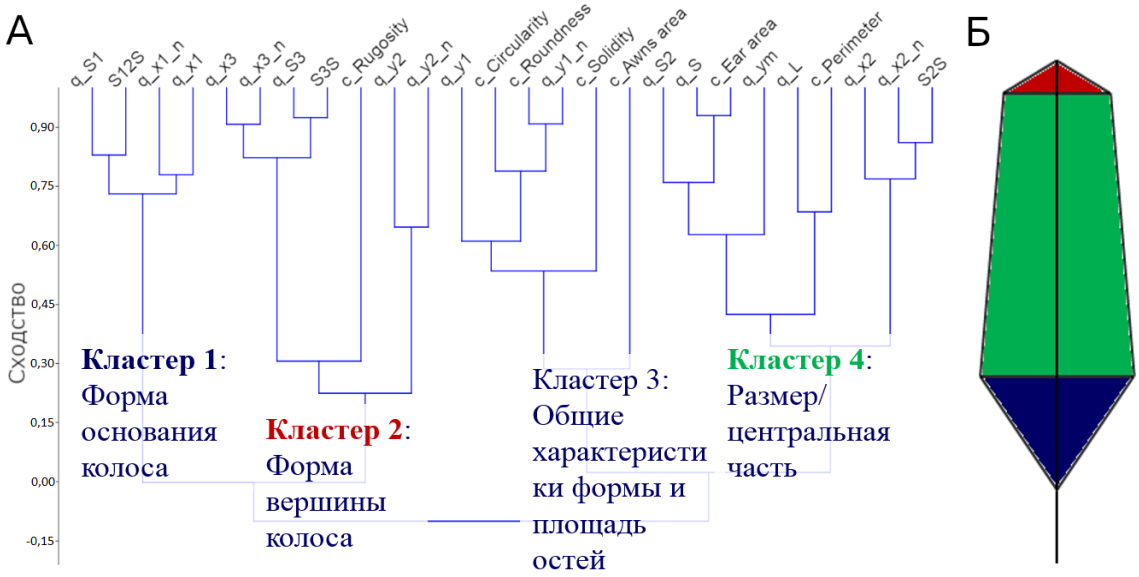


Рисунок 28. А) Результат кластеризации характеристик формы колоса на основе меры близости признаков d_r. Б) Представление формы колоса в виде модели четырехугольников. Синим обозначен сегмент колоса у основания. Зеленым – центральный сегмент колоса. Красным – сегмент колоса у вершины.

На полученном дереве визуализации корреляции можно выделить четыре кластера признаков. Слева направо: первый кластер включает характеристики площади и линейных размеров сегмента основания модели четырехугольников; второй кластер включает характеристики формы вершины колоса – длина вершинного сегмента, его площадь, ширину у вершинного сегмента, а также индекс шероховатости контура колоса; третий кластер включает характеристики формы

контура колоса, такие как округлость (Circularity), закругленность (Roundness), целостность формы (Solidity), площадь остей (Awns area); последний, четвертый кластер включает в себя параметры размера колоса, такие как площадь аппроксимирующих колос четырехугольников (S), площадь контура колоса (Ear area), его периметр (Perimeter) и длину колоса (L).

Таким образом, в предложенном подходе форма колоса характеризуется тремя основными сегментами: основание, центральная часть и вершина. Характеристики, относящиеся к этим сегментам между собой статистически связаны. Группы этих параметров могут отражать как особенности генетического контроля формы колоса, так и специфику модели описания формы колоса. Четвертый же кластер отражает общие параметры формы (вытянутость, округлость и площадь остей).

4.5.2 Анализ вариабельности морфометрических характеристик колоса

Для того, чтобы определить какие из полученных характеристик наиболее значимо характеризуют тот или иной генотип, был проведен анализ корреляции между значениями признаков для колосьев всех генотипов. Для этого был использован анализ главных компонент (PCA).

Как можно видеть из таблицы 12 параметры колоса имеют существенно различающиеся диапазоны значений. Так средняя длина колоса внутри генотипа варьируется от 46,69 мм до 123,18 мм, при среднем стандартном отклонении этого параметра по всем генотипам 17,39 мм.

При этом средняя площадь колоса варьируется от 247,58 мм^2 до 665,37 мм^2, при среднем стандартном отклонении этого параметра по всем генотипам 145,65 мм^2. В то же время индексы формы контура (округлость, закругленность, целостность и шероховатость) варьируются в пределах от 0,05 до 1,39. Данное различие в диапазонах параметров связано с их различной природой. Длина колоса измеряется в миллиметрах, площадь в квадратных миллиметрах, а индексы формы величины безразмерные и отображают относительные характеристики. Другими словами, полученные характеристики являются разномасштабными. Это означает, что использовать матрицу вариаций-ковариаций на данном наборе характеристик не

целесообразно. Более корректно использовать матрицу корреляций, нормирующую значения анализируемых характеристик на значения их дисперсии.

Согласно графику 29, распределение образцов различных видов заметно различаются. Так виды *T. aestivum* и *T. compactum* распределены по всей видимой области, но в большей части *T. aestivum* представлен в левом верхнем квадранте, а *T. compactum* в правом верхнем квадранте. Вид *T. spelta* представлен только в левой части графика, а виды *T. antiquorum* и *T. sphaerococcum* представлены только в нижней части. Зеленые линии на графике обозначают характеристики и их вклад в первые две компоненты, объясняющие дисперсию. Таким образом, сорта различаются прежде всего по размеру колосьев.

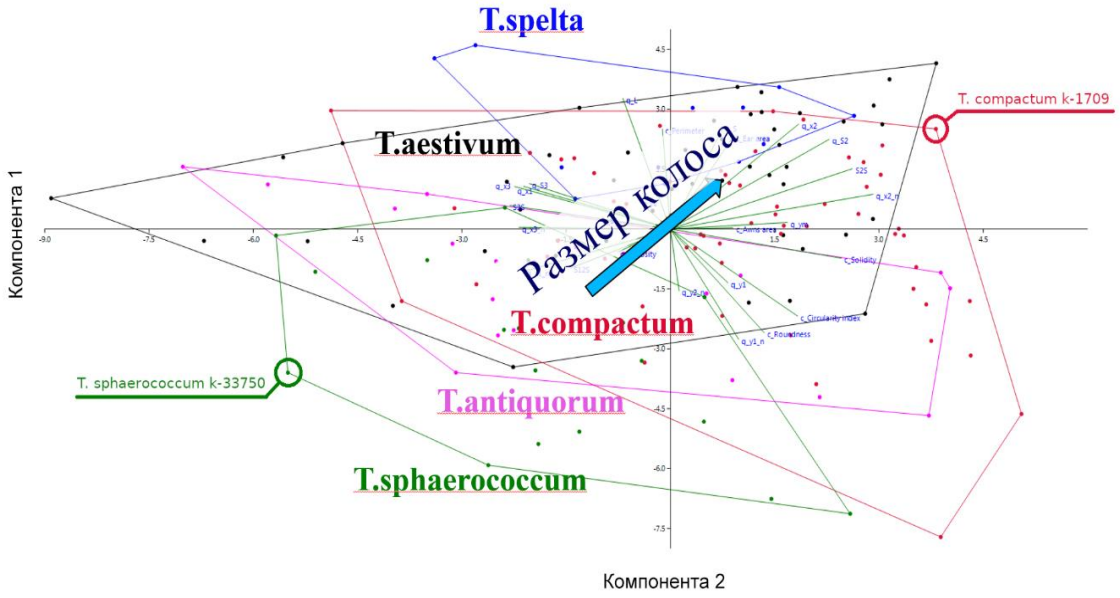


Рисунок 29. Точечный график разложения параметров колосьев на первые две главные компоненты. Черным – образцы вида T. aestivum.; синим – T. spelta; малиновым – T. compactum.; зеленым – T. sphaerococcum; пурпурным - T. antiquorum. Зеленой и красной окружностями выделены два контрастных колоса разных видов.

Ниже на рисунках 30 и 31 представлены диаграммы рассеивания отношения площади и длины колоса из модели четырехугольников и площади контура колоса. Отношение площади контура колоса и площади четырехугольников хорошо согласуется (коэффициент корреляции Пирсона R = 0,93; крит. значение a = 0,11), в то время как у отношения длины и площади контура нет такого соответствия. В

правой нижней части графика распределены образцы *T. spelta*. В левой верхней части образцы *T. compactum*. В нижней левой части *T. sphaerococcum*. Образцы же видов *T. aestivum* и *T. antiquorum* распределились в основном в растянутой области центральной диагонали от левого нижнего (*T. antiquorum*) до правого верхнего угла (*T. aestivum*.) графика.

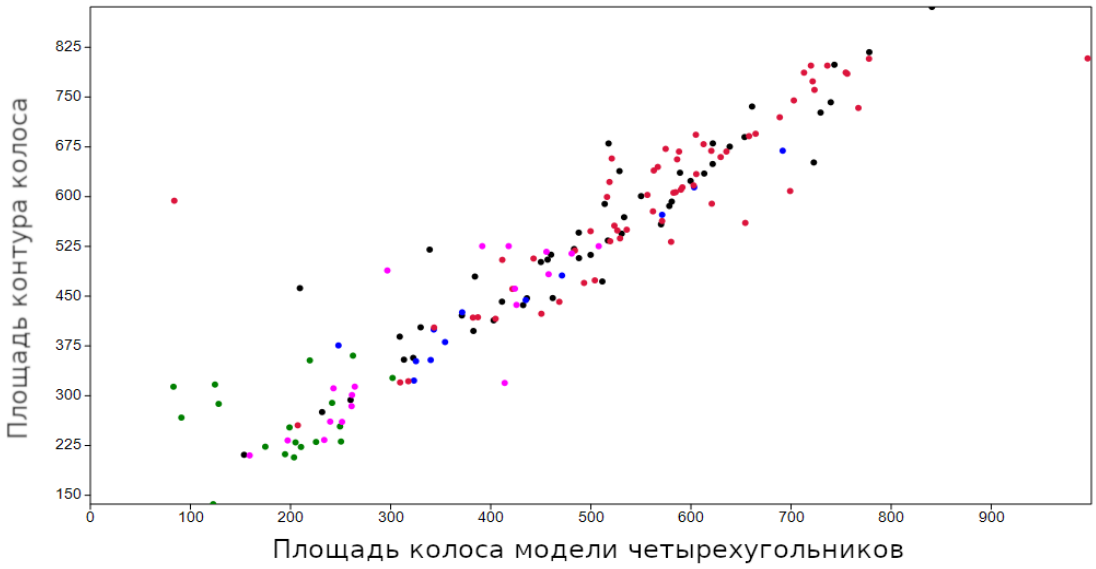


Рисунок 30. Распределение отношения площади контура колоса к площади колоса в модели четырехугольников. Черным – образцы вида T. aestivum.; синим – T. spelta; малиновым – T. compactum.; зеленым – T. sphaerococcum; пурпурным - T. antiquorum.

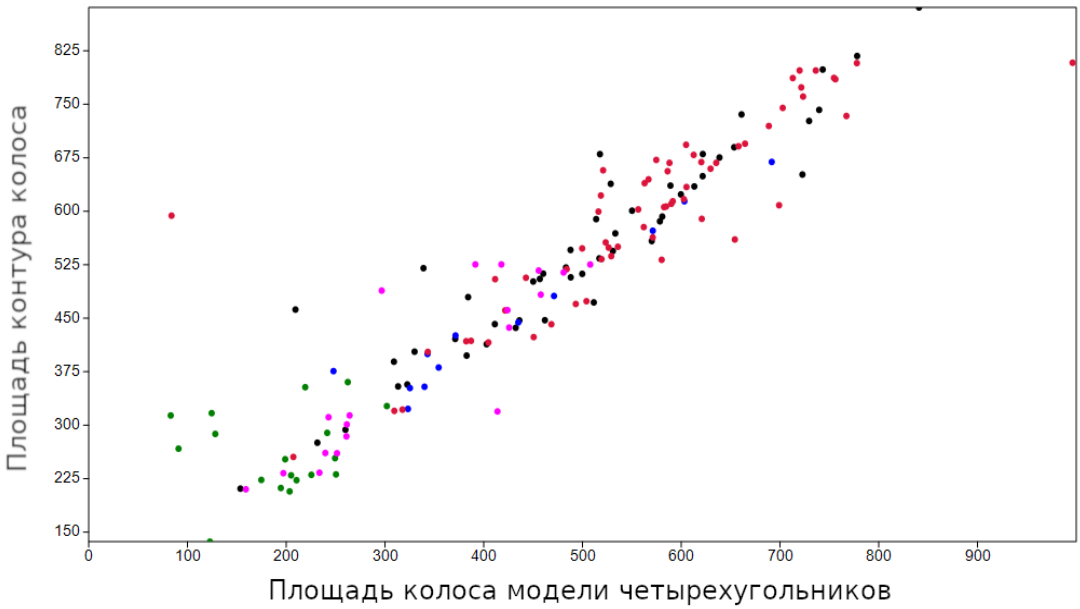


Рисунок 31. Распределение отношения площади колоса в модели четырехугольников к длине колоса. Черным – образцы вида T. aestivum.; синим – T. spelta; малиновым – T. compactum.; зеленым – T. sphaerococcum; пурпурным - T. antiquorum.

Рассмотрим влияние видовой принадлежности на основные вычисляемые характеристики формы колоса. Однофакторный дисперсионный анализ выявил ряд статистически значимо различающихся параметров, приведенных в таблице 13. Значимые характеристики выделены жирным шрифтом.

Таблица 13. Результаты однофакторного дисперсионного анализа характеристик колоса по видовой принадлежности в качестве фактора. Полужирным выделены значимые значения для видовой дискриминации.

Обозначение характеристики	**Название**	**Значение F-статистики**	***p*-value**
c_Perimeter	Периметр колоса	21,39	**4,59E-14**
c_Ear_area	Площадь колоса	31,95	**<2e-16**
c_Awns_area	Площадь остей колоса	26	**<2e-16**
c_Circularity	Индекс округлости контура колоса	12,24	**1,12E-08**
c_Roundness	Индекс закругленности контура колоса	11,1	**6,12E-08**
c_Rugosity	Индекс шероховатости контура колоса	0,682	*0,606*
c_Solidity	Индекс целостности контура колоса	5,603	**0,000307**
q_x1	Высота сегмента основания колоса	0,589	*0,671*
q_x1_n	Нормированная высота сегмента основания колоса	5,472	**0,000379**
q_x2	Высота центрального сегмента колоса	29,99	**<2e-16**
q_x2_n	Нормированная высота	6,286	**0,000103**

Обозначение характеристики	**Название**	**Значение F-статистики**	***p*-value**
	центрального сегмента колоса		
q_x3	Высота вершинного сегмента колоса	2,251	0,066
q_x3_n	Нормированная высота вершинного сегмента колоса	2,279	0,0633
q_y1	Ширина колоса у основания	5,464	**0,000384**
q_y1_n	Нормированная ширина колоса у основания	7,151	**0,0000261**
q_y2	Ширина колоса у вершины	17,2	**1,07E-11**
q_y2_n	Нормированная ширина колоса у вершины	15,91	**6,22E-11**
q_ym	Нормированная площадь	67,74	**<2e-16**
q_L	Длина колоса	27,99	**<2e-16**
q_S1	Площадь сегмента основания колоса	4,654	**0,00142**
q_S2	Площадь центрального сегмента колоса	19,86	**3,25E-13**
q_S3	Площадь вершинного сегмента колоса	0,446	*0,775*
q_S	Площадь модели четырехугольников	30,87	**<2e-16**
S1S	Нормированная площадь сегмента основания колоса	0,963	*0,43*
S2S	Нормированная площадь центрального сегмента колоса	7,946	**0,00000748**

Обозначение характеристики	**Название**	**Значение F-статистики**	***p*-value**
S3S	Нормированная площадь вершинного сегмента колоса	1,78	*0,135*

Основные характеристики, по которым различия значимы связаны с характеристиками формы контура колоса (индексы формы) и его размерами (периметр, площадь колоса, площадь остей), а также линейные размеры формы колоса из модели четырехугольников (длина, ширина колоса).

Чтобы уточнить влияние видовой принадлежности на некоторые отдельные признаки был проведен анализ попарных сравнений по критерию Манна-Уитни с поправкой Бонферрони на множественную проверку гипотез. Анализ показал, что большинство генотипов различаются по таким признакам как: q_x2, q_y2, q_L, q_S2, q_S, что согласуется с однофакторным дисперсионным анализом. Однако виды *T. sphaerococcum* и *T. aestivum* выглядят довольно слабо различающимися: они не значимо различаются по характеристикам q_x2, q_y2, q_L (таблицы 14 и 15), значения для этих пар отмечены красным), которые различаются для большинства остальных пар. Тем не менее, для таких характеристик как q_S и q_S2 эти виды значимо различаются.

Таблица 14. Парные различия (*p*-value) по критерию Манна-Уитни с коррекцией Бонферрони для длины центрального сегмента колоса (q_x2) из модели четырехугольников. Полужирным шрифтом выделены значимые значения для видовой дискриминации.

q_x2	T. *aestivum*	T. *spelta*	T. *compactum*	T. *sphaerococcum*	T. *antiquorum*
T. *aestivum*		**0,02103**	**0,00316**	**1,462E-05**	**0,0004642**
T. *spelta*	**0,02103**		**2,35E-06**	**5,343E-05**	**5,055E-05**
T. *compactum*	**0,00316**	**2,35E-06**		**2,373E-06**	**0,008433**
T. *sphaerococcum*	**1,462E-05**	**5,343E-05**	**2,373E-06**		0,157
T. *antiquorum*	**0,0004642**	**5,055E-05**	**0,008433**	0,157	

Таблица 15. Парные различия (*p*-value) по критерию Манна-Уитни с коррекцией Бонферрони для длины колоса. Полужирным шрифтом выделены значимые значения для видовой дискриминации.

q_L	T. *aestivum*	T. *spelta*	T. *compactum*	T. *sphaerococcum*	T. *antiquorum*
T. *aestivum*		**0,0175**	**2,596E-05**	**1,6E-06**	**0,001038**
T. *spelta*	**0,0175**		**1,166E-05**	**7,963E-05**	**0,0002255**
T. *compactum*	**2,596E-05**	**1,166E-05**		**0,0009168**	0,8307
T. *sphaerococcum*	**1,6E-06**	**7,963E-05**	**0,0009168**		0,3739
T. *antiquorum*	**0,001038**	**0,0002255**	0,8307	0,3739	

4.5.3 Вывод

Полученные параметры позволяют дифференцировать виды, в основном по линейным размерам колоса, в первую очередь по длине. Кластерный анализ показал группировку параметров согласующуюся с предложенной моделью описания формы колоса. Это демонстрирует адекватность предложенного подхода - данные морфометрии колосьев согласуются с интерпретацией получаемых параметров.

Результаты анализа главных компонент показали, что виды *T. aestivum* и *T. compactum*, вид *T. spelta* и виды *T. antiquorum* и *T. sphaerococcum* разделяются по получаемым параметрам.

Дисперсионный анализ выявил ряд характеристик, значимо различающихся для различных пар видов пшеницы. Виды *T. sphaerococcum* и *T. aestivum* оказались слабо различимы по специфическим для большинства других пар видов характеристикам. Однако они оказались различимы для других параметров, что позволяет сделать вывод о возможности дифференцировать их друг от друга на основе комплексных признаков.

4.6 Заключение по главе 4

Разработанный метод морфометрии колоса пшеницы позволяет:

- распознавать образ колоса на изображениях, полученных по двум стандартным протоколам;

- отделять на изображении ости от “тела” колоса;
- оценивать количественные характеристики остистости колоса (площадь остей на изображении, количество и среднюю длину остей);
- оценивать количественные характеристики формы колоса, как интегральные, так и описывающие форму в виде специальной модели.

Анализ точности выделения области колоса показал, что наибольшее влияние на точность определения границ колоса, его “тела” и остей вносит масштаб изображения: чем дальше от объектива камеры расположен колос, тем больше ошибка сегментации. Масштаб съемки существенно различался в зависимости от протокола получения изображений. В протоколе «на столе» камера находится значительно ближе к снимаемому объекту. В протоколе «на прищепке» расстояние до объектов также могло быть различное. При условии, что фокусное расстояние камеры было фиксировано, масштаб зависел только от расстояния до объекта съемки.

Цветокоррекция изображения существенно не влияла на точность сегментации. Вероятнее всего это явилось следствием стандартизованного освещения при съемке колоса. Тем не менее, возможность цветокоррекции заложена в предложенный метод. Ожидается, что она позволит с большей объективностью оценивать цвета на изображениях и поможет существенно улучшить результаты обработки в условиях слабо контролируемого освещения.

Среднее значение J_a меньше по сравнению с J_b отчасти связано с малой площадью остей относительно всего колоса и одновременно с большим количеством граничных пикселей для областей остей. Таким образом, цена ошибки на стадии бинаризации значительно выше для области остей, чем для области “тела” колоса. Кроме того, ручное маркирование изображений высокого разрешения является трудоемким, и не позволяет до конца избежать флуктуации при выделении границ колоса.

Был предложен способ оценки параметров формы колоса. Они условно могут быть разделены на «общие» характеристики, которые отражают форму целиком

(например, округлость, площадь, длина колоса) и параметры модели, описывающей колос в виде двух четырехугольников с общей стороной. Среди значимых для дискриминации видов параметров преобладают общие характеристики колоса и параметры модели четырехугольников. Полученная в результате анализа информация о морфометрических характеристиках колоса может быть применена в высокопроизводительном, автоматизированном фенотипировании и при проведении селекционных экспериментов.

Проведенная оценка точности показала высокую точность распознавания образа колоса для автоматизированного сбора морфометрических данных о колосьях.

Несмотря на невысокую точность в распознавании областей остей колосьев по сравнению с отпечатками, сделанными вручную, данный подход может быть использован для верификации аннотации имеющихся фотографий колосьев.

Сегментация колоса на области остей и “тела” колоса может существенно улучшить оценку формы колоса за счет устранения искажений, вносимых остями в образ колоса.

ГЛАВА 5. ИНФОРМАЦИОННАЯ СИСТЕМА ДЛЯ АННОТАЦИИ МОРФОМЕТРИЧЕСКИХ ХАРАКТЕРИСТИК КОЛОСА ПШЕНИЦЫ

Сбор количественных данных по характеристикам колосьев, полученных экспертами, а также накопление данных морфометрии полученных автоматизированным путем – важный этап разработки методов машинного анализа. Получаемые данные о фенотипе включают широкий спектр характеристик. Данные разнородны и накапливаются в большом объеме. Для их накопления, хранения и систематизации необходима разработка базы данных.

Для сбора, хранения и анализа информации о фенотипических (морфометрических) характеристиках колоса пшеницы была разработана компьютерная информационная система SpikeDroid. Она позволяет систематизировать и аннотировать накопленные на основе анализа изображений данные. Система состоит из базы данных (SpikeDroidDB), web-интерфейса на основе CMS (CMF) Drupal и модулей обработки изображений, позволяющих производить автоматизированную аннотацию загруженных в базу данных изображений. База данных обеспечивает структурированное хранение цифровых изображения колоса и их аннотаций, а также предоставляет пользователю гибкую систему запросов для доступа к данным. Web-интерфейс системы доступен по адресу http://spikedroid.biores.cytogen.ru и позволяет работать с ней, как со стационарных компьютеров, так и с мобильных устройств. SpikeDroid имеет модуль импорта и экспорта данных и изображений.

5.1 Модель данных

Характеристики колоса, которые были использованы для описания его фенотипа, приведены в таблице 16. Нужно учитывать, что фенотип растения формируется на основе генотипа под влиянием окружающей среды. В системе SpikeDroid был выделен ряд понятий важных при проведении селекционно-генетических исследований и определены отношения между ними. Логическая модель данных включает таблицу “растения”, связанную с тремя блоками

информации – коллекцией, окружающей средой и фенотипом колоса. Текущая версия базы данных содержит 5 таблиц и 4 отношения между ними (рисунок 32).

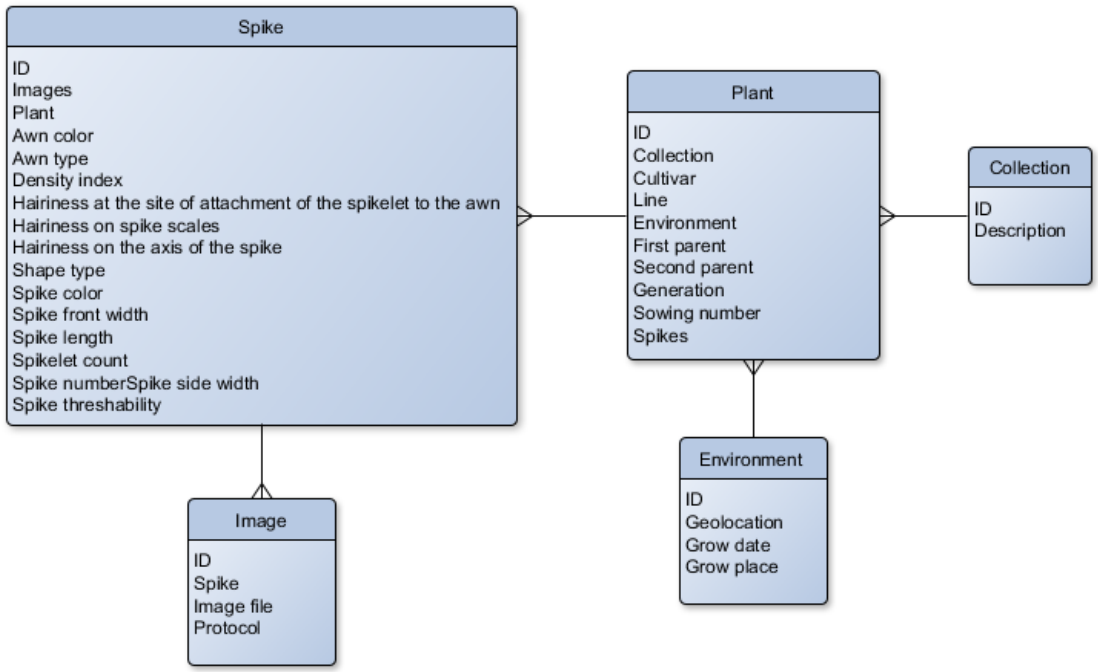


Рисунок 32. Блок-схема базы данных SpikeDroidDB. (ER-модель)

Таблица Plant (Растение) включает описание сорта (образца) растения или линии, ссылки на родительские растения, поколение и посевной номер растения. Растение связано с таблицами Collection (Коллекция), Environment (место произрастания) и Spike (описание колоса). Для коллекции указывается её держатель и аффилиация (описание). Место произрастания содержит геоданные, дату вегетации, фенологические данные и данные об условиях выращивания растения. Фенотип колоса описывается набором из 14 признаков, представленных в таблице 16. Нами была проанализирована онтология пшеницы портала CropOntology.org и выполнена привязка признаков, содержащийся в нашей базе данных с признаками онтологии CropOntology.org, в тех случаях где это представлялось возможным.

Таблица 16. Перечень характеристик, описывающих фенотип колоса в базе данных SpikeDroidDB.

Наименование атрибута	**Наименование атрибута (eng)**	**Значения**	**Термин онтологии CropOntology.org**
Номер колоса	Spike number	Первый номер присваивается главному колосу.	
Длина колоса	Spike length	Единица измерения: см	CO_321:0000056

Ширина лицевой стороны колоса	Spike front width	Единица измерения: см	
Ширина боковой стороны колоса	Spike side width	Единица измерения: см	
Количество колосков	Spikelet count	Единица измерения: шт	CO_321:0000058
Плотность колосков	Density index	D= [(A-1) × 10] / B Где: (A-1) – число колосков в колосе без верхушечного колоска; B – длина стержня колоса в см.	CO_321:0000055
Цвет колоса	Spike color	красный/черный/белый	
Опушение на колосковых чешуях	Hairiness on spike scales	есть/нет (выбор из двух вариантов)	
Опушение на оси колоса	Hairiness on the axis of the spike	есть/нет (выбор из двух вариантов)	
Опушение на месте прикрепления колоска к оси	Hairiness at the site of attachment of the spikelet to the axis	есть/нет (выбор из двух вариантов)	
Тип остей.	Awn type	Безостый, остистый (длина остей длиннее или равна чем длина колоса), короткоостистый (длина остей короче, чем длина колоса), полуостистый (верх длина остей длиннее чем вниз длина остей)	CO_321:0000027
Цвет остей	Awn color	красный/белый/черный/фиолетовый/янтарный/смешанный	CO_321:0000960
Ломкоколосость	Spike threshability	ломкий/не ломкий (выбор из двух вариантов)	CO_321:0000659
Форма колоса	Shape type	спельта/норма/компактная	

5.2 Технологии реализации системы SpikeDroid

Система SpikeDroid была разработана на основе Drupal 8 (Abbott and Jones, 2016). Для ее разработки были использовали следующие модули:

- “Conditional fields” – позволяет задавать правила появления/сокрытия полей в зависимости от установленных условий;
- “Display suit” – расширяет возможности представлений материалов контента Drupal;
- “Field_group” – позволяет группировать поля для управления ими как единым целым;
- “Geolocation” – добавляет новый тип полей-геолокаций, что позволяет использовать Google Places API;
- “Charts” – модуль для создания диаграмм и графиков.

Web-интерфейс системы SpikeDroid разработан при помощи технологии адаптивной верстки. Это позволят работать с системой, используя различные типы устройств (мобильные телефоны, персональные или планшетные компьютеры).

5.3 Модуль интерфейса системы SpikeDroid

Система SpikeDroid доступна по адресу http://spikedroid.biores.cytogen.ru. Она содержит краткую информацию о базе данных, ссылки для входа в систему или регистрации и ссылки на основные блоки информации в базе данных. Пользователь может получить доступ к базе данных, зарегистрировавшись на сайте. Зарегистрированный пользователь имеет возможность добавлять и аннотировать собственные растения. Для того чтобы просмотреть список растений, информация о которых доступна в базе, необходимо с главной страницы перейти по ссылке «Plants». После этого осуществляется переход на страницу списка растений (рисунок 33).

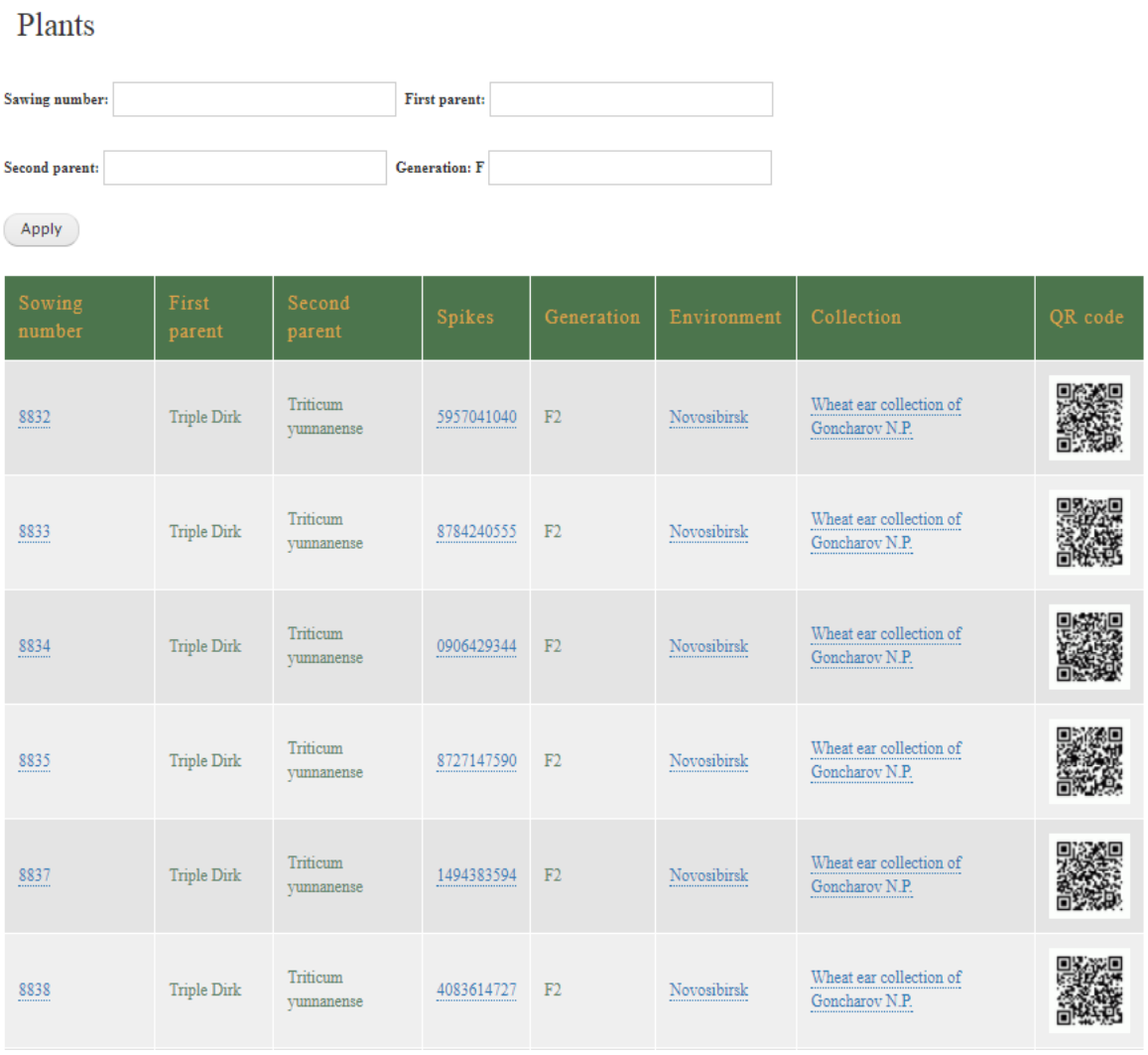


Рисунок 33. Фрагмент страницы со списком образцов растений представленным в базе данных SpikeDroidDB.

На этой странице отображается информация о растениях в виде таблицы. Информация различного типа (фенотип колоса, информация о месте произрастания, генотипе и коллекции) располагается в разных колонках таблицы и доступна по соответствующим гиперссылкам. В крайней правой колонке для каждого растения приводятся ссылки на его QR-код (от английского Quick Response Code). QR-код является матричным (двумерным) штрих-кодом, который может быть сканирован камерой мобильного устройства. Информация, которую он содержит, может содержать до 4296 символов цифр и букв, что достаточно для описания доступа к растению в нашей базе по протоколу HTTP. QR-код присваивается в нашей базе каждому растению, может быть распечатан на плотной бумаге и физически прикреплен к растению. После этого для получения или изменения параметров растения достаточно считать этот код мобильным устройством. Таким образом, интерфейс SpikeDroid позволяет сохранять информацию об измерениях фенотипов растений в ходе эксперимента непосредственно в базу данных, минуя записи в полевых и лабораторных журналах исследователей.

На странице описания растения (рисунок 34) отображается информация о генотипе растения, данные о месте произрастания, которые отображаются с помощью метки на карте. Помимо этого, на странице приводятся ссылки на описание фенотипа колоса и коллекции, а также отображается QR-код для доступа к странице растения.

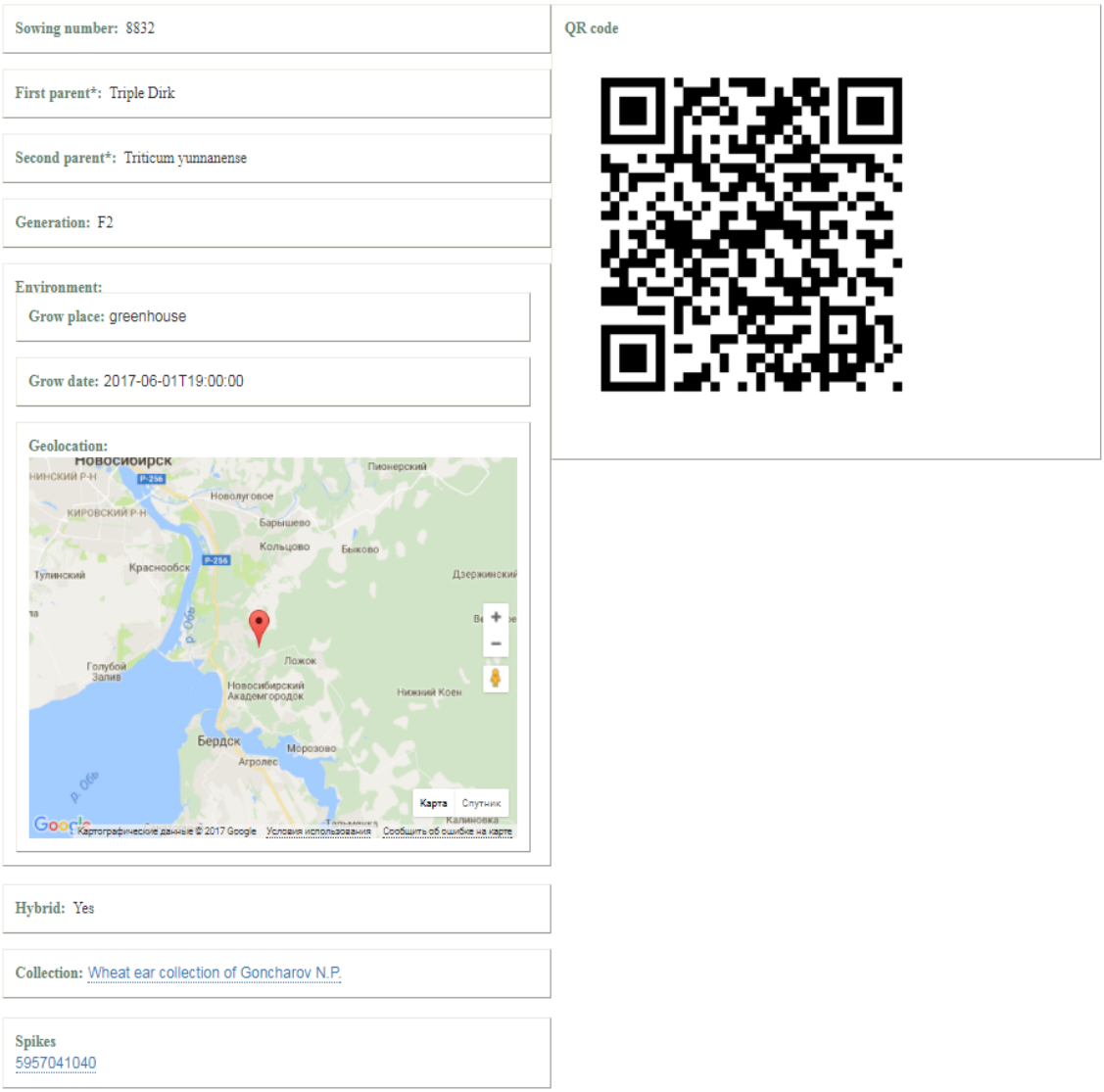


Рисунок 34. Страница описания растения. Раздел содержит информацию о коллекции, посевном номере растения, родителях, поколении, месте и условиях взращивания (поле/теплица).

На странице описания фенотипа колоса отображаются характеристики колоса (рисунок 35). Список всех возможных характеристик представлен в таблице 16. С каждым колосом можно соотнести набор фотографий. Для каждой фотографии указывается протокол, с помощью которого она была получена.

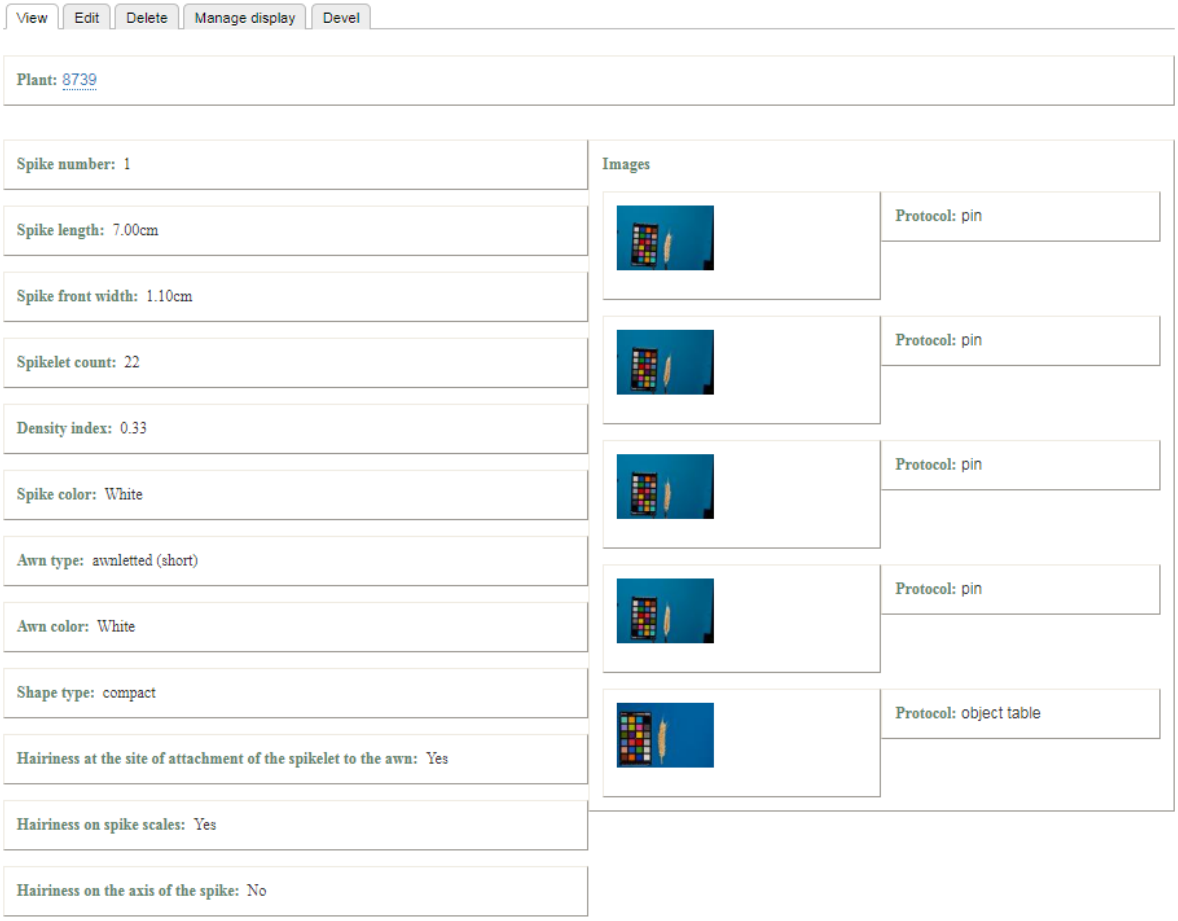


Рисунок 35. Страница описания фенотипа колоса. Раздел содержит данные экспертного фенотипирования колоса пшеницы и краткую аннотацию его изображений.

5.4 Информационное содержание базы данных SpikeDroidDB

Текущая версия базы данных содержит коллекцию колосьев F2 гибридов от скрещивания австралийского сорта мягкой пшеницы Triple Dirk на образец KU506 китайской пшеницы *T. yunnanense*. Коллекция включает в себя 380 растений и 1475 фотографий колосьев. Аннотирование колосьев осуществлялось экспертом по списку морфологических характеристик, представленных в таблице 16. В таблице 17 представлены результаты разделения колосьев по некоторым признакам. На рисунке 36 показано распределение колосьев по форме, длине и ширине.

Таблица 17. Разделение колосьев F2 гибридов от скрещивания австралийского сорта мягкой пшеницы Triple Dirk на образец KU506 китайской пшеницы *Triticum yunnanense* по форме колоса, остистости, опушению колосовой чешуи, опушению на оси колоса, опушению на месте прикрепления колоса к оси.

Форма колоса	**Компакт**	**Норма**	**Спельта**
Количество, шт	6	35	63
Остистость	безостный	короткоостистый	

Количество, шт	80	24	
Опушение колосовой чешуи	присутствует	отсутствует	
Количество, шт	68	36	
Опушение на оси колоса	присутствует	отсутствует	
Количество, шт	93	11	
Опушение на месте прикрепления колоса к оси	присутствует	отсутствует	
Количество, шт	74	30	

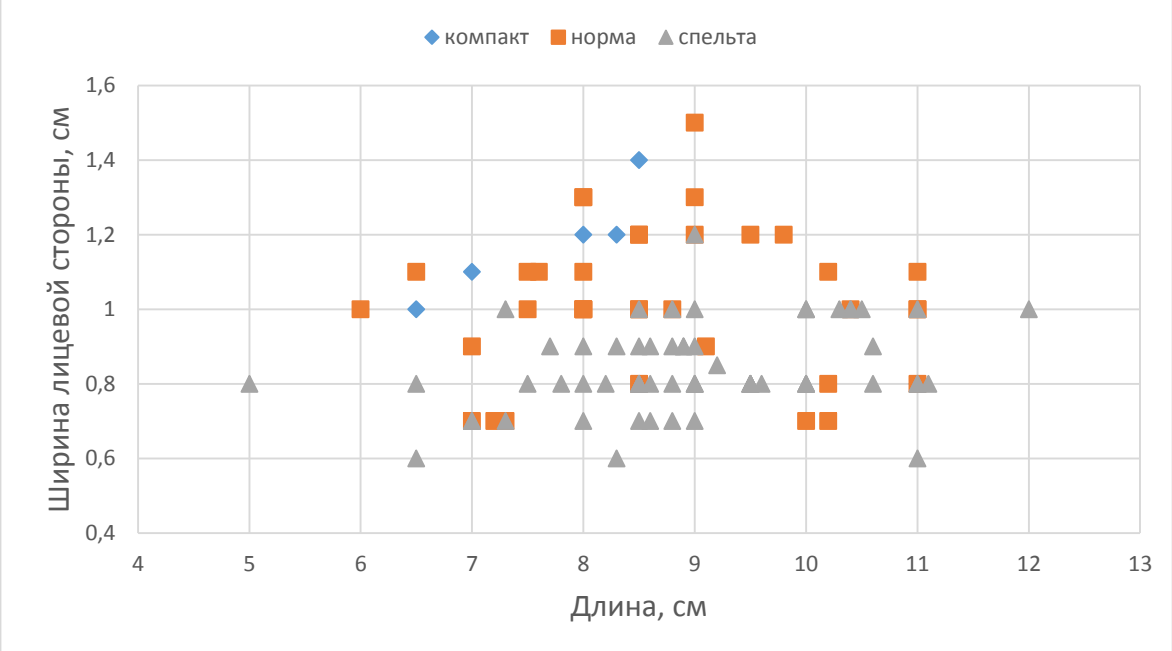


Рисунок 36. Распределение колосьев гибридов поколения F2 от скрещивания австралийского сорта мягкой пшеницы Triple Dirk на образец KU506 китайской пшеницы *T. yunnanense* по длине и ширине.

5.5 Заключение по главе 5

Система SpikeDroid разработана для сбора, хранения и анализа информации о фенотипических характеристиках колоса пшеницы. SpikeDroid позволяет структурированно хранить цифровые изображения колоса, производить их аннотацию и предоставляет пользователю гибкую систему запросов для доступа к данным. В системе SpikeDroid был выделен ряд понятий, важных при проведении селекционно-генетических исследований, включающих генотип, фенотип, условия выращивания, и определили отношения между ними. Так, например, фенотип колоса описывается набором из 14 признаков. Используя эту систему, была произведена оцифровка и аннотация коллекции колосьев F2 гибридов от скрещивания австралийского сорта мягкой пшеницы Triple Dirk на образец KU506 китайской

пшеницы *T. yunnanense*. Проведен анализ изменчивости колосьев по форме, длине и ширине.

Web-интерфейс системы SpikeDroid доступен по адресу http://spikedroid.biores.cytogen.ru и позволяет работать с системой, как со стационарных компьютеров, так и с мобильных устройств.

ЗАКЛЮЧЕНИЕ

Обзор современной литературы по генетике и селекции пшеницы показал, что одним из основных направлений исследований является создание сортов и линий с высокой продуктивностью. Это обуславливает активное изучение репродуктивных органов растений, колосьев и зерен, определения механизмов генетического контроля их размеров и формы. Для успешного решения этих задач необходимо проводить массовые оценки фенотипических характеристик репродуктивных органов пшеницы. Несмотря на то, что в современных селекционно-генетических экспериментах участвуют десятки-тысяч растений, фенотипирование колосьев осуществляется преимущественно вручную, что является трудоемкой задачей, а также влечет за собой субъективизм, присущий человеку. Стоит отметить также, что результаты измерений в этих экспериментах документируются вручную без использования современных информационных технологий. Создание высокопроизводительных технологий фенотипирования колосьев и зерен злаков позволит существенно упростить и ускорить проведение селекционных экспериментов и обеспечит структурирование и систематизацию полученных данных.

Настоящая работа посвящена созданию и апробации методов автоматизированного фенотипирования зерен и колосьев пшеницы на основе анализа двухмерных цветных цифровых изображений.

В работе предложен новый метод определения количественных характеристик зерен пшеницы на основе анализа изображений с использованием мобильных устройств. Оценка точности метода показала, что ошибки в подсчете количества зерен составляют порядка 2%, а в оценке размеров зерен порядка 8%, что является приемлемым для решения задачи оценки количества зерен в колосе (типичное количество зерен не превышает 50-60 шт.).

Метод является достаточно быстрым и может быть использован для высокопроизводительного фенотипирования, в том числе и в условиях за пределами лаборатории. Мы показали, что использование разработанного метода позволяет производить дифференцировку различных сортов по размеру и форме зерен.

Мы предложили новый метод определения количественных характеристик колоса пшеницы на основе анализа двухмерных изображений. Метод включает два протокола для получения изображений, пригодных для последующего эффективного анализа, а также реализованное консольное приложение WERecognizer, позволяющее распознавать колосья пшеницы на полученных изображениях и извлекать из них морфометрические характеристики колосьев. Предложенный метод показал высокую точность при оценке площади колоса и его остей на изображении.

На примере 14 генотипов пяти видов гексаплоидных пшениц показано, что получаемые предложенным методом характеристики колосьев распределяются на четыре кластера: один связанный с общими характеристиками формы колоса, и три с отдельными сегментами колоса – основанием, серединой и вершиной.

Для информационной поддержки экспериментов по анализу колоса пшеницы создана компьютерная система SpikeDroid, которая интегрирует данные из различных источников: от пользователя, с анализируемых изображений, из сети Интернет. Разработанная система позволила собрать информацию о более чем 2000 растений, провести аннотацию и автоматизированный компьютерный анализ изображений их колосьев.

ВЫВОДЫ

1. Разработано приложение SeedCounter для мобильных устройств и персональных компьютеров, позволяющее оценивать с высокой точностью количество и размеры зерен пшеницы, расположенных на листе белой бумаги стандартного формата.
2. Впервые разработан метод определения количественных характеристик размера, формы и остистости колоса пшеницы на основе анализа двухмерных изображений, реализованный в виде приложения WERecognizer для персональных компьютеров.
3. Впервые предложена геометрическая модель представления колоса в виде двух четырехугольников, позволяющая оценить его форму и размер. Модель характеризует основные количественные параметры трех сегментов колоса: основания, центральной части и вершины.
4. Анализ 14 генотипов пяти видов гексаплоидных пшениц с помощью предложенной модели позволил выявить признаки колоса, значимо различающиеся для отдельных генотипов: размер центральной части, длину колоса, ширину основания, площадь центрального сегмента и основания колоса. Полученные данные полностью согласуются с существующей классификацией и подтверждают её.
5. Разработана база данных SpikeDroidDB для накопления, хранения, систематизации и поиска информации о фенотипических признаках колоса. Система SpikeDroid позволяет автоматически анализировать изображения колосьев пшеницы, полученных по установленному протоколу, содержит данные экспертной аннотации колоса пшеницы; морфометрические характеристики колоса и зерен; изображения колосьев; данные о образцах растений и их генотипе; описания коллекций образцов, месте их произрастания.

СПИСОК ЛИТЕРАТУРЫ

1. Брагина М. К., Афонников Д. А., Салина Е. А. Прогресс в секвенировании геномов растений–направления исследований // Вавиловский журнал генетики и селекции. – 2019. – Т. 23. – №. 1. – С. 38-48.
2. Генаев М. А., Дорошков А. В., Пшеничникова Т. А., Морозова Е. В., Симонов А. В., и др. Информационная поддержка селекционно-генетического эксперимента у пшеницы в системе WheatPGE // Математическая биология и биоинформатика. – 2012. – Т. 7. – №. 2. – С. 410-424.
3. Гончаров Н. П. Определитель разновидностей мягкой и твердой пшениц. – 2009.
4. Гончаров Н. П. Сравнительная генетика пшениц и их сородичей. – Общество с ограниченной ответственностью Академическое издательство Гео, - 2012.
5. Гультяева Е. И., Левитин М. М., Семенякина Н. Ф., Никифорова Н. В., Савельева Н. И. Болезни зерновых культур в Северо-Западном регионе России // Защита и карантин растений. – 2007. – №. 6.
6. Дьяконов В. П. Matlab 6/6.1/6. 5+ Simulink 4/5. Основы программирования: руководство пользователя // М.: Солон-Пресс. – 2002.
7. Ильясова Н. Ю., Куприянов А. В., Храмов А. Г. Информационные технологии анализа изображений в задачах медицинской диагностики // М.: Радио и связь. – 2012. – С. 424.
8. Кошкин Е. И., Панфилова О. Ф., Пильщикова Н. В. Частная физиология полевых культур. – 2005.
9. Местецкий Л. М. Математические методы распознавания образов. – 2008
10. Моргун В. В., Швартау В. В., Киризий Д. А. Физиологические основы формирования высокой продуктивности зерновых злаков // Физиология и биохимия культурных растений. – 2010.
11. Подколодный Н.Л., Подколодная О. А. Онтологии в биоинформатике и системной биологии // Вавиловский журнал генетики и селекции. – 2016. – Т. 19. – №. 6. – С. 652-660.

12. Санин С. С., Мотовилин А. А., Корнева Л. Г., Жохова Т. П., Полякова Т. М., и др. Химическая защита пшеницы от болезней при интенсивном зернопроизводстве // Защита и карантин растений. – 2011. – №. 8.
13. Степанов С. А., Горюнов А. А., Кузьмина А. В., Агапова А. В. Лимитирующие эндогенные факторы продуктивности яровой пшеницы // Бюллетень Ботанического сада Саратовского государственного университета. – 2008. – №. 7.
14. Фляксбергер К. А. Определитель настоящих хлебов. – Alexander Doweld, 1922.
15. Фляксбергер К. А. Пшеницы. – Гос. изд-во колхозной и совхозной лит-ры, 1935. – Т. 1.
16. Форсайт Д., Понс Ж. Компьютерное зрение. Современный подход //М.: Вильямс. – 2004. – Т. 928. – С. 22.
17. Халафян А. А. Statistica 6. Математическая статистика с элементами теории вероятностей. – 2010.
18. Abbott N., Jones R. Learning Drupal 8. – Packt Publishing Ltd, 2016.
19. Alaux M., Letellier T., Alfama-Depauw F., Jamilloux V., Rogers J., et al. IWGSC Sequence Repository: Moving towards tools to facilitate data integration for the reference sequence of wheat // PAG XXIV-Plant and Animal Genome Conference. – 2016.
20. Appels R., Eversole K., Stein N., Feuillet C., Keller B., et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome // Science. – 2018. – V. 361. – №. 6403.
21. Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., et al. Gene Ontology: tool for the unification of biology // Nat genet. – 2000. – V. 25. – №. 1. – P. 25-9.
22. Avraham S., Tung C. W., Ilic K., Jaiswal P., Kellogg E. A., et al. The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations // Nucleic acids research. – 2008. – V. 36. – №. suppl_1. – P. D449-D454.
23. Bai X. D., Cao Z. G., Wang Y., Yu Z. H., Zhang X. F., et al. Crop segmentation from images by morphology modeling in the CIE L* a* b* color space // Computers and electronics in agri-culture. – 2013. – V. 99. – P. 21-34.

24. Benfey P.N., Mitchell-Olds T. From genotype to phenotype: systems bi-ology meets natural variation // Science. – 2008. – V. 320. – №. 5875. – P. 495-497.
25. Bennett M.D., Leitch I.J. Nuclear DNA amounts in angiosperms—583 new estimates // Annals of Botany. – 1997. – V. 80. – №. 2. – P. 169-196.
26. Berry J. C., Fahlgren N., Pokorny A. A., Bart R. S., Veley K. M. An automated, high-throughput method for standardizing image color profiles to improve image-based plant phenotyping // PeerJ. – 2018. – V. 6. – P. e5727.
27. Bi K., Jiang P., Li L., Shi B., Wang C. Non-destructive measurement of wheat spike characteristics based on morphological image processing // Transactions of the Chinese Society of Agricultural Engineering. – 2010. – V. 2010. – №. 12.
28. Braun H. J., Atlin G., Payne T. Multi-location testing as a tool to identify plant response to global climate change // Climate change and crop production. – 2010. – V. 1. – P. 115-138.
29. Busemeyer L., Mentrup D., Möller K., Wunder E., Alheit K., et al. Breedvision — A multi-sensor platform for non-destructive field-based phenotyping in plant breeding // Sensors. – 2013. – V. 13. – №. 3. – P. 2830-2847.
30. Canny J. A computational approach to edge detection // IEEE Transactions on pattern analysis and machine intelligence. – 1986. – №. 6. – P. 679-698.
31. Chalupska D., Lee H. Y., Faris J. D., Evrard A., Chalhoub B., et al. Acc homoeoloci and the evolution of wheat genomes // Proceedings of the National Academy of Sciences. – 2008. – V. 105. – №. 28. – P. 9691-9696.
32. Chen X., Xun Y., Li W., Zhang J. Combining discriminant analysis and neural networks for corn variety identification // Computers and electronics in agriculture. – 2010. – V. 71. – P. S48-S53.
33. Darrigues A., Hall J., van der Knaap E., Francis D. M., Dujmovic N., et al. Tomato analyzer-color test: a new tool for efficient digital phenotyping // Journal of the American Society for Horticultural Science. – 2008. – V. 133. – №. 4. – P. 579-586.
34. Dawson-Howe K. A practical introduction to computer vision with opencv. – John Wiley & Sons, 2014.
35. DeWitt N., Guedira M., Lauer E., Sarinelli M., Tyagi P., et al. Sequence-based mapping identifies a candidate transcription repressor underlying awn suppression at the B1 locus in wheat // New Phytologist. – 2020. – V. 225. – №. 1. – P. 326-339.

36. Dorofeev V. F., Filatenko A. A., Migushova E. F., Udachin R. A., Jakubciner M. M. Wheat, flora of cultivated plants, Vol. 1 // Kolos. – 1979. – V.1.
37. Duan L., Yang W., Huang C., Liu Q. A novel machine-vision-based facility for the automatic evaluation of yield-related traits in rice // Plant Methods. – 2011. – V. 7. – №. 1. – P. 44.
38. Eliceiri K. W., Berthold M. R., Goldberg I. G., Ibáñez L., Manjunath B. S., et al. Biological imag-ing software tools // Nature methods. – 2012. – V. 9. – №. 7. – P. 697.
39. Ercan S., Ertugrul F., Aydin Y., Akfirat F. S., Hasancebi S., et al. An EST-SSR marker linked with yellow rust resistance in wheat // Biologia Plantarum. – 2010. – V. 54. – №. 4. – P. 691-696.
40. Evers T., Millar S. Cereal grain structure and development: some implications for quality //Journal of cereal science. – 2002. – V. 36. – №. 3. – P. 261-284.
41. FAO, 2011: Crop Prospects and Food Situation. Food and Agriculture Organization, Global Information and Early Warning System, Trade and Markets Division (EST), Rome, Italy.,- 2011.
42. Farooq S., Shahid M., Khan M. B., Hussain M., Farooq M. Improving the productivity of bread wheat by good management practices under terminal drought // Journal of Agronomy and Crop Science. – 2015. – V. 201. – №. 3. – P. 173-188.
43. Filippa G., Cremonese E., Migliavacca M., Galvagno M., Forkel M., G. et al. Phenopix: AR package for image-based vegetation phenology // Agricultural and Forest Meteorology. – 2016. – V. 220. – P. 141-150.
44. Flicek P., Amode M. R., Barrell D., Beal K., Brent S., et al. Ensembl 2012 //Nucleic acids research. – 2012. – V. 40. – №. D1. – P. D84-D90.
45. Furbank R. T., Tester M. Phenomics–technologies to relieve the phenotyping bottleneck // Trends in plant science. – 2011. – V. 16. – №. 12. – P. 635-644.
46. Gale M. D., Devos K. M. Comparative genetics in the grasses // Proceedings of the National Academy of Sciences. – 1998. – V. 95. – №. 5. – P. 1971-1974.
47. Gegas V. C., Nazari A., Griffiths S., Simmonds J., Fish L., et al. A genetic framework for grain size and shape variation in wheat // The Plant Cell. – 2010. – V. 22. – №. 4. – P. 1046-1056.

48. Genaev M. A., Doroshkov A. V., Pshenichnikova T. A., Kolchanov N. A., Afonnikov D. A. Extraction of quantitative characteristics describing wheat leaf pubescence with a novel image-processing technique // Planta. – 2012. – V. 236. – №. 6. – P. 1943-1954.
49. Genaev M. A., Komyshev E. G., Smirnov N. V., Kruchinina Y. V., Goncharov N. P., et al. Morphometry of the wheat spike by analyzing 2D images // Agronomy. – 2019. – V. 9. – №. 7. – P. 390.
50. Gene Ontology Consortium et al. The gene ontology: enhance-ments for 2011 //Nucleic acids research. – 2011. – C. gkr1028.
51. Golzarian M. R., Frick R. A., Rajendran K., Berger B., Roy S., et al. Accurate inference of shoot biomass from high-throughput images of cereal plants // Plant methods. – 2011. – V. 7. – №. 1. – P. 2.
52. Gómez-Robledo L., López-Ruiz N., Melgosa M., Palma A. J., Capitán-Vallvey L. F., et al. Using the mobile phone as Munsell soil-colour sensor: An experiment under controlled illumination conditions // Computers and electronics in agriculture. – 2013. – V. 99. – P. 200-208.
53. Goncharov N. P. Comparative genetic study of tetraploid forms of common wheat without D genome // Генетика. – 1997. – V. 33. – №. 5. – P. 660-663.
54. Goncharov N. P. Comparative genetics of wheats and their related species // Siberian un-ty press, Novosibirsk (Russian). – 2002.
55. Grady L. Random walks for image segmentation // IEEE transac-tions on pat-tern analysis and machine intelligence. – 2006. – V. 28. – №. 11. – P. 1768-1783
56. Grady L., Schwartz E. L. Isoperimetric graph partitioning for im-age segmenta-tion // IEEE transactions on pattern analysis and machine intelligence. – 2006. – V. 28. – №. 3. – P. 469-475.
57. Granitto P. M., Verdes P. F., Ceccatto H. A. Large-scale investigation of weed seed identification by machine vision // Computers and Elec-tronics in Ag-riculture. – 2005. – V. 47. – №. 1. – P. 15-24.
58. Guo Z., Chen D., Alqudah A. M., Röder M. S., Ganal M. W., et al. Genome-wide association analyses of 54 traits identified multiple loci for the determination of floret fertility in wheat // New Phytologist. – 2017. – V. 214. – №. 1. – P. 257-270.

59. Guyon I., Weston J., Barnhill S., Vapnik V. Gene selection for cancer classification using support vector machines // Machine learning. – 2002. – V. 46. – №. 1-3. – P. 389-422.
60. Hancock J. M. (ed.). Phenomics. – CRC Press, 2014.
61. Haque M. A., Takayama A., Watanabe N., Kuboyama T. Cytological and genetic mapping of the gene for four-awned phenotype in Triticum carthlicum Nevski // Genetic Resources and Crop Evolution. – 2011. – V. 58. – №. 7. – P. 1087-1093.
62. Hasançebi S., Mert Z., Ertugrul F., Akan K., Aydin Y., et al. An EST-SSR marker, bu099658, and its potential use in breeding for yellow rust resistance in wheat // Czech Journal of Genetics and Plant Breeding. – 2014. – V. 50. – №. 1. – P. 11-18.
63. Herridge R. P., Day R. C., Baldwin S., Macknight R. C. Rapid analysis of seed size in Arabidopsis for mutant and QTL discovery // Plant methods. – 2011. – V. 7. – №. 1. – P. 1-11.
64. Herridge R.P., Day R.C., Baldwin S., Macknight R.C. Rapid anal-ysis of seed size in Arabidopsis for mutant and QTL discovery // Plant methods. – 2011. – V. 7. – №. 1. – P. 1-11.
65. Houde M., Belcaid M., Ouellet F., Danyluk J., Monroy A. F., et al. Wheat EST resources for functional genomics of abiotic stress // BMC genomics. – 2006. – V. 7. – №. 1. – P. 149.
66. Howse J. Android application programming with OpenCV 3. – Packt Publishing Ltd, 2015.
67. Huang D., Zheng Q., Melchkart T., Bekkaoui Y., Konkin D. J., et al. Dominant inhibition of awn development by a putative zinc-finger transcriptional repressor expressed at the B1 locus in wheat // New Phytologist. – 2020. – V. 225. – №. 1. – P. 340-355.
68. Ilic K., Kellogg E. A., Jaiswal P., Zapata F., Stevens P. F., et al. The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant // Plant physiology. – 2007. – V. 143. – №. 2. – P. 587-599.
69. International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome // Science. – 2014. – V. 345. – №. 6194. – P. 1251788.

70. Jaccard P. The distribution of the flora in the alpine zone. 1 // New phytologist. – 1912. – V. 11. – №. 2. – P. 37-50.
71. Jahnke S., Roussel J., Hombach T., Kochs J., Fischbach A., et al. phenoSeeder-A robot system for automated handling and phenotyping of individual seeds // Plant physiology. – 2016. – V. 172. – №. 3. – P. 1358-1370.
72. Jantasuriyarat C., Vales M. I., Watson C. J. W., Riera-Lizarazu O. Identification and mapping of genetic loci affecting the free-threshing habit and spike compactness in wheat (Triticum aestivum L.) // Theoretical and Applied Genetics. – 2004. – V. 108. – №. 2. – P. 261-273.
73. Jia J., Zhao S., Kong X., Li Y., Zhao G., et al. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation // Nature. – 2013. – V. 496. – №. 7443. – P. 91.
74. Kaehler A., Bradski G. Learning OpenCV 3: computer vision in C++ with the OpenCV library. – " O'Reilly Media, Inc.", 2016.
75. Karakas O., Gurel F., Uncuoglu A. A. Exploiting a wheat EST database to assess genetic diversity // Genetics and molecular biology. – 2010. – V. 33. – №. 4. – P. 719-730.
76. Katabuchi M. LeafArea: an R package for rapid digital image analysis of leaf area // Ecological Research. – 2015. – V. 30. – №. 6. – P. 1073-1077.
77. Kerber E. R., Rowland G. G. Origin of the free threshing character in hexaploid wheat // Canadian Journal of Genetics and Cytology. – 1974. – V. 16. – №. 1. – P. 145-154.
78. Komyshev E., Genaev M., Afonnikov D. Evaluation of the SeedCounter, a mobile application for grain phenotyping // Frontiers in plant science. – 2017. – V. 7. – P. 1990.
79. Konopatskaia I., Vavilova V., Blinov A., Goncharov N. P. Spike morphology genes in wheat species (Triticum L.) // Proceedings of the Latvian Academy of Sciences. Section B. Natural, Exact, and Applied Sciences. – Sciendo, 2016. – V. 70. – №. 6. – P. 345-355.
80. Kuhl F. P., Giardina C. R. Elliptic Fourier features of a closed contour // Computer graphics and image processing. – 1982. – V. 18. – №. 3. – P. 236-258.

81. Kumar N., Belhumeur P.N., Biswas A., Jacobs D.W., Kress W.J., et al. Leafsnap: A computer vision system for automatic plant species identification // European conference on computer vision. – Springer, Berlin, Heidelberg, 2012. – P. 502-516.
82. Li L., Zhang Q., Huang D. A review of imaging techniques for plant phenotyping // Sensors. – 2014. – V. 14. – №. 11. – P. 20078-20111.
83. Ling H. Q., Zhao S., Liu D., Wang J., Sun H., et al. Draft genome of the wheat A-genome progenitor Triticum urartu // Nature. – 2013. – V. 496. – №. 7443. – P. 87.
84. Long S. P., Ort D. R. More than taking the heat: crops and global change // Current opinion in plant biology. – 2010. – V. 13. – №. 3. – P. 240-247.
85. Ma L., Li T., Hao C., Wang Y., Chen X., et al. Ta GS 5-3A, a grain size gene selected during wheat improvement for larger kernel and yield // Plant biotechnology journal. – 2016. – V. 14. – №. 5. – P. 1269-1280.
86. Madin J., Bowers S., Schildhauer M., Krivov S., Pennington D., et al. An ontology for describing and synthesizing ecological observation data // Ecological informatics. – 2007. – V. 2. – №. 3. – P. 279-296.
87. Manske G. G. B., Ortiz-Monasterio J. I., Van Ginkel M., Gonzalez R. M., Fischer R. A., et al. Importance of P uptake efficiency versus P utilization for wheat yield in acid and calcareous soils in Mexico // European Journal of Agronomy. – 2001. – V. 14. – №. 4. – P. 261-274.
88. Marx V. Biology: The big challenges of big data // Nature. – 2013. – V. 498. – №. 7453. – P. 255-260.
89. Matteis L., Chibon P. Y., Espinosa H., Skofic M., Finkers H. J., et al. Crop ontology: vocabulary for crop-related concepts. – 2013.
90. Meinshausen N., Bühlmann P. Stability selection // Journal of the Royal Statistical Society: Series B (Statistical Methodology). – 2010. – V. 72. – №. 4. – P. 417-473.
91. Muqaddasi Q. H., Brassac J., Koppolu R., Plieske J., Ganal M. W., et al. TaAPO-A1, an ortholog of rice ABERRANT PANICLE ORGANIZATION 1, is associated with total spikelet number per spike in elite European hexaploid winter wheat (Triticum aestivum L.) varieties // Scientific reports. – 2019. – V. 9. – №. 1
92. Novaro P., Colucci F., Venora G., D'egidio M. G. Image analysis of whole grains: a noninvasive method to predict semolina yield in durum wheat // Cereal chemistry. – 2001. – V. 78. – №. 3. – P. 217-221.

93. O'Driscoll A., Daugelaite J., Sleator R. D. 'Big data', Hadoop and cloud computing in genomics // Journal of biomedical informatics. – 2013. – V. 46. – №. 5. – P. 774-781.
94. Paproki A., Sirault X., Berry S., Furbank R., Fripp J. A novel mesh processing based technique for 3D plant analysis // BMC plant biology. – 2012. – V. 12. – №. 1. – P. 1-13.
95. Pau G., Fuchs F., Sklyar O., Boutros M., Huber W. EBImage—an R package for image processing with applications to cellular phenotypes // Bioinformatics. – 2010. – V. 26. – №. 7. – P. 979-981.
96. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., et al. Scikit-learn: Machine learning in Python // Journal of machine learning research. – 2011. – V. 12. – P. 2825-2830.
97. Pethybridge S. J., Nelson S. C. Leaf Doctor: A new portable application for quantifying plant disease severity // Plant disease. – 2015. – V. 99. – №. 10. – P. 1310-1316.
98. Petrie H. Review of STATISTICA 6.0 // British Journal of Mathematical & Statistical Psychology. – 2002. – V. 55. – P. 391.
99. Pourreza A., Pourreza H., Abbaspour-Fard M. H., Sadrnia H. Identification of nine Iranian wheat seed varieties by tex-tural analysis with image processing // Computers and Electronics in Agriculture. – 2012. – V. 83. – P. 102-108.
100. Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P. Numerical recipes in Fortran 77: the art of scientific computing. – Cambridge : Cambridge university press, 1992. – V. 2. – P. 915.
101. Quintana J., Garcia R., Neumann L. A novel method for color correction in epiluminescence microscopy // Computerized Medical Imaging and Graphics. – 2011. – V. 35. – №. 7-8. – P. 646-652.
102. Rahaman M., Chen D., Gillani Z., Klukas C., Chen M. Advanced phenotyping and phenotype data analysis for the study of plant growth and development // Frontiers in plant science. – 2015. – V. 6.
103. Rao M. V. P. Telocentric mapping of the awn inhibitor gene Hd on chromosome 4B of common wheat // Cereal Research Communications. – 1981. – P. 335-337.

104. Reshef D. N., Reshef Y. A., Finucane H. K., Grossman S. R., McVean G., et al. Detecting novel associations in large data sets // science. – 2011. – V. 334. – №. 6062. – P. 1518-1524.

105. Roerdink J. B. T. M., Meijster A. The watershed transform: Definitions, algorithms and parallelization strategies // Fundamenta informaticae. – 2000. – V. 41. – №. 1, 2. – P. 187-228.

106. Roussel J., Geiger F., Fischbach A., Jahnke S., Scharr H. 3D surface reconstruction of plant seeds by volume carving: performance and accuracies // Frontiers in plant science. – 2016. – V. 7. – P. 745.

107. Sears E. R. Chromosome mapping with the aid of telocentrics // Proc. 2nd Intern. Wheat Genet. Symp.. Hereditas Suppl. – 1966. – V. 2. – P. 370-381.

108. Shapiro L. G., Stockman G. C. Computer Vision, ch. 12. – 2001.

109. Shi J., Malik J. Normalized cuts and image segmentation // IEEE Transactions on pattern analysis and machine intelligence. – 2000. – V. 22. – №. 8. – P. 888-905

110. Shrestha R., Matteis L., Skofic M., Portugal A., McLaren G., et al. Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontol-ogy developed by the crop communities of practice // Frontiers in physi-ology. – 2012. – V. 3.

111. Simonetti M. C., Bellomo M. P., Laghetti G., Perrino P., Simeone R., et al. Quantitative trait loci influencing free-threshing habit in tetraploid wheats // Genetic Resources and Crop Evolution. – 1999. – V. 46. – №. 3. – P. 267-271.

112. Simons K. J., Fellers J. P., Trick H. N., Zhang Z., Tai Y. S., et al. Molecular characterization of the major wheat domestication gene Q // Genetics. – 2006. – V. 172. – №. 1. – P. 547-555.

113. Smith C. L., Eppig J. T. The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data // Mammalian genome. – 2012. – V. 23. – №. 9-10. – P. 653-668.

114. Sood S., Kuraparthy V., Bai G., Gill B. S. The major threshability genes soft glume (sog) and tenacious glume (Tg), of diploid and polyploid wheat, trace their origin to independent mutations at non-orthologous loci // Theoretical and Applied Genetics. – 2009. – V. 119. – №. 2. – P. 341-351.

115. Sticklen M. B. Feedstock crop genetic engineering for alcohol fuels // Crop science. – 2007. – V. 47. – №. 6. – P. 2238-2248.

116. Strange H., Zwiggelaar R., Sturrock C., Mooney, S.J., Doonan J.H. Automatic estimation of wheat grain morphometry from computed tomography data // Functional Plant Biology. – 2015. – V. 42. – №. 5. – P. 452-459.

117. Suzuki S. Topological structural analysis of digitized binary images by border following // Computer vision, graphics, and image processing. – 1985. – V. 30. – №. 1. – P. 32-46.

118. Swaminathan M. S., Rao M. V. P. Macro-mutations and sub-specific differentiation in Triticum. – 1961.

119. Tahir A.R., Neethirajan S., Jayas D.S., Shahin M.A., Symons S.J., et al. Evaluation of the effect of moisture content on cereal grains by digital image analysis // Food Research International. – 2007. – V. 40. – №. 9. – P. 1140-1145

120. Tanabata T., Shibaya T., Hori K., Ebana K., Yano M. Smart-Grain: high-throughput phenotyping software for measuring seed shape through im-age analysis // Plant physiology. – 2012. – V. 160. – №. 4. – P. 1871-1880.

121. Tanabata T., Yamada T., Shimizu Y., Shinozaki Y., Kanekatsu M., et al. Development of automatic segmentation software for efficient measurement of area on the digital images of plant organs // Horticultural Research (Japan). – 2010. – V. 9. – №. 4. – P. 501-506.

122. Team R. C. R: A language and environment for statistical computing. – 2013.

123. Varshney R.K., Nayak S.N., May G.D., Jackson S.A. Next-generation sequencing technologies and their implications for crop genet-ics and breeding // Trends in biotechnology. – 2009. – V. 27. – №. 9. – P. 522-530.

124. Wang D., Yu K., Jin D., Sun L., Chu J., et al. ALI-1, candidate gene of B1 locus, is associated with awn length and grain weight in common wheat // bioRxiv. – 2019. – P. 688085.

125. Whan A. P., Smith A. B., Cavanagh C. R., Ral J. P. F., Shaw L. M., et al. GrainScan: a low cost, fast method for grain size and colour measurements // Plant methods. – 2014. – V. 10. – №. 1. – P. 1-10.

126. Whitley D. A genetic algorithm tutorial // Statistics and computing. – 1994. – V. 4. – №. 2. – P. 65-85.

127. Wiesnerová D., Wiesner I. Computer image analysis of seed shape and seed color for flax cultivar description // Computers and elec-tronics in agriculture. – 2008. – V. 61. – №. 2. – P. 126-135.

128. Wu Z., Leahy R. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation // IEEE transactions on pattern analysis and machine intelligence. – 1993. – V. 15. – №. 11. – P. 1101-1113.

129. Yazdanbakhsh N., Fisahn J. High throughput phenotyping of root growth dynamics, lateral root formation, root architecture and root hair development enabled by PlaRoM // Functional Plant Biology. – 2009. – V. 36. – №. 11. – P. 938-946.

130. Yu J. K., Dake T. M., Singh S., Benscher D., Li W., et al. Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat // Genome. – 2004. – V. 47. – №. 5. – P. 805-818.

131. Zahn C. T. Graph-theoretical methods for detecting and describing gestalt clus-ters // IEEE Transactions on computers. – 1971. – V. 100. – №. 1. – P. 68-86.

132. Zapotoczny P. Discrimination of wheat grain varieties using image analysis and neural networks. Part I. Single kernel texture // Journal of Cereal Science. – 2011. – V. 54. – №. 1. – P. 60-68.

133. Zayas I., Pomeranz Y., Lai F. S. Discrimination of wheat and nonwheat components in grain samples by image analysis // Cereal Chemistry. – 1989. – V. 66. – №. 3. – P. 233-237

134. Zeybek A., Yigit F. Determination of virulence genes frequencies in wheat stripe rust (Puccinia striiformis f. sp. tritici) populations during natural epidemics in the regions of southern Aegean and western Mediterranean in Turkey // Pak J Biol Sci. – 2004

135. Zhang Z., Belcram H., Gornicki P., Charles M., Just J., et al. Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat // Proceedings of the National Academy of Sciences. – 2011. – V. 108. – №. 46. – P. 18737-18742.