Пермский филиал федерального государственного автономного образовательного учреждения высшего образования   
«Национальный исследовательский университет   
«Высшая школа экономики»

Факультет экономики, менеджмента и бизнес-информатики

*Комлев Савелий Алексеевич*

***Разработка информационной системы для автоматизации филогенетического исследования геномов***

*Курсовая работа*

по направлению подготовки *38.03.05 Бизнес-информатика*

образовательная программа бакалавриата «Бизнес-информатика»

Руководитель:

Преподаватель кафедры ИТ в бизнесе

*Найданов Илья Валерьевич*

Пермь, 2022 год

Оглавление

[Оглавление 2](#_Toc99486616)

[Введение 4](#_Toc99486617)

[Глоссарий 6](#_Toc99486618)

[Глава 1. Анализ предметной области 7](#_Toc99486619)

[1. Филогенетика 7](#_Toc99486620)

[2. Филогенетическое исследование геномов 9](#_Toc99486621)

[3. Структурное моделирование бизнес-процессов 13](#_Toc99486622)

[3.1 Контекстная диаграмма 13](#_Toc99486623)

[3.2 Уровень А0: филогенетическое исследование генмов 13](#_Toc99486624)

[3.3 Уровень А1: подготовка входных данных 14](#_Toc99486625)

[3.4 Уровень А2: выделение и обработка кластеров ортологичных генов 14](#_Toc99486626)

[3.5 Уровень А3: подготовка данных для отчётов 15](#_Toc99486627)

[3.6 Уровень А4: создание отчётов 15](#_Toc99486628)

[3.7 Выводы из структурного моделирования 15](#_Toc99486629)

[4. Детализация структурных диаграмм 16](#_Toc99486630)

[5. Требования к ИС 18](#_Toc99486631)

[6. Требования к данным 20](#_Toc99486632)

[6.1 Требования к входным данным 20](#_Toc99486633)

[6.2 Требования к выходным данным 21](#_Toc99486634)

[Глава 2. Проектирование 22](#_Toc99486635)

[1. Проектирование архитектуры ИС 22](#_Toc99486636)

[1.1 Диаграмма компонент 22](#_Toc99486637)

[1.2 Диаграмма компонент в расширенной нотации 23](#_Toc99486638)

[2. Проектирование процессов ИС 24](#_Toc99486639)

[2.1 Процесс подготовки входных данных 24](#_Toc99486640)

[2.2 Процесс выделения и обработки кластеров ортологичных генов 24](#_Toc99486641)

[2.3 Процесс подготовки данных для отчётов 25](#_Toc99486642)

[2.4 Процесс создания отчётов 25](#_Toc99486643)

[Глава 3. Разработка и тестирование 26](#_Toc99486644)

[1. Принципы разработки конвейеров для анализа данных 26](#_Toc99486645)

[2. Разработка конвейера для анализа данных 27](#_Toc99486646)

[3. Тестирование 27](#_Toc99486647)

[3.1 Запуск программы 27](#_Toc99486648)

[3.2 Отчёт с филогенетическим деревом 29](#_Toc99486649)

[3.3 Гены под отбором 30](#_Toc99486650)

[Заключение 31](#_Toc99486651)

[Библиографический список 32](#_Toc99486652)

[ПРИЛОЖЕНИЕ A 33](#_Toc99486653)

Введение

Биоинформатика — междисциплинарная область, объединяющая общую биологию, молекулярную биологию, кибернетику, генетику, химию, компьютерные науки, математику и статистику. Последние 20 лет на фоне развития информационных и биологических технологий эта область очень бурно развивалась и двигала вперёд не только фундаментальную науку, но и индустрию: от досконального изучения эволюционной истории самых разных видов организмов до персонализированного подбора противораковой терапии на основе генетических данных пациента. Среди прочего биоинформатика перевернула филогенетику, область биологической систематики, которая занимается выявлением и прояснением эволюционных взаимоотношений среди разных видов жизни на Земле, сделав её наукой о больших молекулярных данных.

Несмотря на многочисленные преимущества, идея использовать биоинформатические методы в филогенетике привела к появлению существенной проблемы в виде существования множества узкоспециальных программ, разработанных не связанными друг с другом коллективами учёных, и результаты работы которых необходимо объединять для получения содержательных выводов. Узкая специализация программ связана со сложностью решаемых в филогенетике задач: каждая программа хорошо решает свою небольшую, но сложную задачу. Для проведения полноценного исследования необходимо последовательно запустить несколько программ, подавая на вход каждой последующей программы файлы, полученные предыдущими программами. В ряде случаев передаваемые между программами файлы необходимо дополнительно преобразовывать вручную, поскольку формат выхода и входа разных программ не всегда согласован.

Предлагаемое решение описанной выше проблемы: разработать информационную систему (ИС), осуществляющую последовательный запуск специальных программ и промежуточное преобразование файлов автоматически в соответствии с паттерном проектирования "[Конвейер](https://en.wikipedia.org/wiki/Pipeline_(software))". Такая ИС будет значительно экономить время исследователей и сокращать количество допускаемых ошибок при запуске программ и преобразовании файлов.

Актуальность разработки ИС для автоматизации филогенетического исследования геномов обусловлена потребностью исследователей в существовании такой системы. На момент написания этой работы мне не удалось найти в открытом доступе систему, позволяющую полностью автоматизировать филогенетическое исследование геномов.

Объект исследования – процесс филогенетического исследования геномов.

Предмет исследования – информационная система для автоматизации процесса филогенетического исследования геномов.

Цель – разработка информационной системы для автоматизации процесса филогенетического исследования геномов.

Задачи:

* Провести анализ процесса филогенетического исследования геномов;
* сформировать требования к ИС для автоматизации филогенетического исследования геномов;
* спроектировать ИС для автоматизации филогенетического исследования геномов;
* реализовать и протестировать приложение.

Для реализации приложения будут использоваться язык программирования Python, интегрированная среда разработки программного обеспечения PyCharm, система управления рабочими процессами и инструмент для создания воспроизводимых и масштабируемых конвейеров для анализа данных Snakemake.

Глоссарий

* Геном — совокупность [наследственного](https://ru.wikipedia.org/wiki/%D0%9D%D0%B0%D1%81%D0%BB%D0%B5%D0%B4%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D1%81%D1%82%D1%8C) материала, заключённого в [клетке](https://ru.wikipedia.org/wiki/%D0%9A%D0%BB%D0%B5%D1%82%D0%BA%D0%B0) организма. В рамках системы будет представлен в виде текстового файла.
* Аннотация генома — это определение кодирующих и регуляторных последовательностей ДНК в геноме. Например, белковых генов. Выполняется с помощью специальных программ.
* Последовательность гена — описание химического состава гена в виде текстовой последовательности на четырёхбуквенном алфавите, где каждая буква соответствует одному нуклеотиду.
* Последовательность белка — описание химического состава белка в виде текстовой последовательности на двадцати-буквенном алфавите, где каждая буква соответствует одной аминокислоте.
* Кластеризация белков (белковых последовательностей) — задача группировки множества последовательностей белков на подмножества (кластеры) таким образом, чтобы последовательности из одного кластера были более похожи друг на друга, чем на последовательности из других кластеров. В результате решения этой задачи определяются кластеры ортологичных групп белков (КОГ).
* Ортологичные гены (белки) — гены (белки), которые в ходе эволюции произошли от общего предка. Гены (белки) считаются ортологичными, если их последовательности похожи на статистически значимом уровне. Поиск ортологичных генов – важная задача эволюционной биологии.
* Выравнивание последовательностей — биоинформатический метод, основанный на размещении двух или более последовательностей мономеров ДНК, РНК или белков друг под другом таким образом, чтобы легко увидеть сходные участки в этих последовательностях. Сходство первичных структур двух молекул может отражать их функциональные, структурные или эволюционные взаимосвязи [7].

Глава 1. Анализ предметной области

1. Филогенетика

Филогенетика — область биологической систематики, которая занимается выявлением и прояснением эволюционных взаимоотношений среди разных видов жизни на Земле, как современных, так и вымерших. Эволюционная теория утверждает, что сходство тех или иных особей или видов часто указывает на общее происхождение или общего предка. Потому взаимоотношения, установленные филогенетической систематикой, часто описывают эволюционную историю видов и их филогенез, исторические взаимоотношения между ветвями организмов или их частей, например, их генов [1].

До второй половины 20-го века различные виды живых организмов сравнивали по морфологии, или по внешней форме организмов, и их внутренней структуре. Эти методы позволили исследователям приблизительно очертить эволюционную историю большинства видов, однако многие вопросы по-прежнему оставались без ответа.

Вместе с бурным развитием методов молекулярной биологии и информационных технологий во второй половине 20-го века филогенетика сильно преобразилась: стало возможным сравнивать организмы на основе структуры молекул ДНК, РНК и белков. Такое сравнение позволяет делать намного более сложные и точные, чем на основе морфологии, поскольку сопоставляются молекулы, несущие полную информацию о строении и функционировании клеток организма. ДНК – это «чертёж», согласно которому каждая клетка в организме проходит весь свой жизненный цикл. Отличия в морфологии – лишь следствие отличий в молекулах ДНК.

В 80-90-х годах благодаря прорывной технологии секвенирования нового поколения ученые по всему миру получили возможность определять последовательности не только отдельных генов или белков, но и целых геномов любого организма. Вскоре количество геномов различных организмов в открытых базах данных начало расти экспоненциально, как показано на рисунке 1 [2].

Chart

Description automatically generatedНа основе анализа больших геномных данных разных видов организмов было сделано множество поразительных открытий – эволюционная биология в целом и филогенетика в частности в корне изменились навсегда. Например, стало известно, что ДНК любого человека совпадает на 99,9% с ДНК с другого человека, на 96% с ДНК шимпанзе, на 85% c ДНК мыши, и на 60% с ДНК банана. Также стало понятно, что есть определенный набор генов, общий для всех организмов на земле, или что грибы генетически ближе к животным, чем к растениям [3]. Подобных удивительных открытий существует великое множество, однако всех их объединяет одна вещь – все они были получены в результате обработки больших данных с помощью специальных программ, автоматизирующих вычисления.

Рисунок 1 – Динамика количества публичных геномных данных с 1985 по 2015

Область исследований, посвященная разработке программ для автоматического анализа биологических данных, называется «биоинформатика» или «вычислительная биология». В качестве отдельной науки биоинформатика существует лишь около 30–40 лет, однако за это время она успела произвести петабайты данных разного типа и тысячи программ для их обработки. Из-за большой сложности разработки и, порой, недостатка финансирования многие из этих программ узкоспециализированы: они выполняют относительно небольшую задачу, часто недостаточную для содержательных биологических выводов, и очень требовательны к входным данным. Кроме того, программы, не снискавшие большой славы, коих большинство, нередко забрасываются разработчиками и не получают не только обновлений для соответствия новым требованиям, но даже исправления багов.

Для того, чтобы решить описанные выше проблемы, в области биоинформатики очень распространена практика объединения нескольких узкоспециализированных программ в одну систему, которая автоматически соединяет выходы одних программ со входами других, попутно дополнительно преобразовывая данные. На выходе такая система выдаёт отчеты, по которым можно делать содержательные выводы. Подобная схема объединения программ называется конвейер (англ. pipeline). Подробнее об этом паттерне проектирования будет рассказано в главе 3.

Прежде, чем приступить к разработке конвейера для автоматизации филогенетического исследования генома, необходимо разобраться с тем, как подобное исследования обычно проводится «вручную».

1. Филогенетическое исследование геномов

Graphical user interface

Description automatically generatedПрежде, чем приступить к детальному описанию процесса филогенетического анализа геномов, рассмотрим диаграмму ключевых понятий выбранной предметной области, изображённую на рисунке 2.

Рисунок 2 - Диаграмма понятий филогенетики

Диаграмма понятий позволяет описать сущности предметной области через свойства сущностей и отношения наследования и агрегации между сущностями. Построенная диаграмма может пригодиться на этапе проектирования ИС.

Филогенетический анализ геномов можно проводить по-разному, однако в большинстве случаев такой анализ включает в себя следующую последовательность шагов [4].

На первом этапе геномы необходимо аннотировать, то есть выделить в их последовательностях важные функциональные элементы, например гены. На последующих этапах исследования именно эти элементы из разных геномов будут подвергнуты анализу и сравнению, поскольку естественный отбор действует по большей части именно на них. Полученные в ходе аннотации гены следует перевести из нуклеотидного алфавита в аминокислотный, то есть получить из последовательностей генов последовательности белков, с которыми смогут работать программы на следующем этапе.

На втором этапе нужно кластеризовать полученные белки, то есть сгруппировать последовательности белков таким образом, чтобы последовательности из одной группы были более похожи друг на друга, чем на последовательности из других групп. При этом, последовательности, не похожие ни на какие другие, отбрасываются. Полученные группы называются кластерами ортологичных групп белков (КОГ) [5]. С биологической точки зрения смысл такого кластера заключается в том, что внутри него оказываются белки, которые предположительно были унаследованы рассматриваемыми видами от общего предка, но их последовательности разошлись в ходе эволюции, причём чем больше различий между последовательностями двух белков, тем раньше они разошлись, или тем более интенсивному отбору они подвергались [6].

Далее нужно объединить результаты аннотации и кластеризации, чтобы получить кластеры ортологичных генов, в которых каждый ген соответствует определённому белку из кластеров белков. То есть мы обратно переводим белки в гены. Это необходимо для того, чтобы в дальнейшем определить, какие гены подвергались отбору на основе анализа мутаций.

Table

Description automatically generated with medium confidenceНа четвёртом этапе отобранные ранее гены из кластеров нужно выровнять, то есть сопоставить последовательности таким образом, чтобы максимизировать суммарное сходство по всем позициям. Пример выравнивания аминокислотных последовательностей изображён на рисунке 3. На нём сопоставлены 7 последовательностей, одинаковые буквы выделены одним и тем же цветом.

Рисунок 3 - Пример выравнивания последовательностей

Выравнивание позволяет понять, в каких позициях последовательностей произошли мутации, и как следствие предположить их эволюционную близость, а также определить позиции под отбором.

Diagram

Description automatically generatedНа пятом этапе следует объединить выравнивания всех рассмотренных ранее генов по каждому организму и на основе них построить филогенетическое дерево, отражающее эволюционные взаимосвязи между различными видами или другими сущностями, имеющими общего предка. Пример такого дерева изображён на рисунке 4.

Рисунок 4 - Пример филогенетического дерева

На последнем этапе нужно на основе выравнивания последовательностей и филогенетического дерева определить, какие гены находятся под отбором. Для этого необходимо для каждого гена каждого рассматриваемого организма рассчитать соотношение количества не синонимичных замен к количеству синонимичных замен по сравнению с аналогичными генами других организмов. Не синонимичными заменами являются замены одного символа на другой в нуклеотидной последовательности, которые не привели к замене символов в соответствующей ей белковой последовательности. Считается, что не синонимичные замены являются нейтральными мутациями, поскольку не меняют структуру белка, а следовательно - не меняют его молекулярные свойства. Синонимичные замены в гене, напротив, приводят к замене и в белке, а значит могут повлиять на его свойства и, следовательно, функции, которые он выполняет в клетке. Таким образом, на основе рассчитанных соотношений можно делать следующие выводы о типе отбора, действующем на исследуемые гены [6]:

* Если соотношение близко к 1, значит количество не синонимичных замен примерно равно количеству не синонимичных, то есть распределение мутаций равномерно по всей последовательности – естественный отбор не благоволит никаким мутациям – все они нейтральные;
* если соотношение заметно больше 1, значит не синонимичных замен больше, чем синонимичных, следовательно закрепились мутации, меняющие структуру белка, а значит на ген действует положительный, закрепляющий отбор;
* если соотношение заметно меньше 1, значит не синонимичных замен меньше, чем синонимичных, следовательно не закрепились мутации, меняющие структуру белка, а значит на ген действует отрицательный, отчищающий отбор отбор.

Для выполнения каждого из описанных выше шагов филогенетического исследования геномов существуют различные программы. Выбор программ для конкретного исследования зависит от специфики данных, с которыми нужно работать. Например, аннотация геномов бактерий сильно отличается от аннотации геномов растений, поскольку их гены имеют очень разную структуру, следовательно программы должны искать разные паттерны в последовательностях. Поскольку разрабатываемый в рамках этой работы инструмент планируется как универсальный, в конвейер будут включены наиболее универсальные и распространенные программы [4], описанные в таблице 1.

Таблица 1 – Программы используемые для каждого этапа исследования.

|  |  |
| --- | --- |
| **Этап исследования** | **Используемая программа** |
| Аннотация геномов | Prokka |
| Кластеризация белков | ProteinOrtho |
| Объединение результатов аннотации и кластеризации | GetFasta и оригинальный код |
| Выравнивание | PRANK |
| Построение филогенетического дерева | IQ-Tree |
| Определение генов под отбором | PAML |

Далее смоделируем процессы филогенетического исследования геномов в виде структурных диаграмм в нотации IDEF0. В диаграммах будут фигурировать описанные выше программы.

1. Структурное моделирование бизнес-процессов

Основой для моделирования послужил детальный текст статьи о масштабном филогенетическом исследовании геномов различных видов морских выдр [4].

* 1. Diagram

     Description automatically generatedКонтекстная диаграмма
  2. Diagram

     Description automatically generatedУровень А0: филогенетическое исследование генмов
  3. Diagram

     Description automatically generatedУровень А1: подготовка входных данных
  4. Diagram

     Description automatically generatedУровень А2: выделение и обработка кластеров ортологичных генов
  5. Diagram

     Description automatically generatedУровень А3: подготовка данных для отчётов
  6. Diagram

     Description automatically generatedУровень А4: создание отчётов
  7. Выводы из структурного моделирования

Внимательно изучив построенные диаграммы, можно обнаружить, что в каждой операции задействован исследователь, который запускает специальные программы, подавая им на вход нужные данные, либо производит промежуточные преобразования данных между запусками программ. Подобная организация процессов: многократный ручной запуск программ через командную строку, ручное преобразование большого количества данных приводит к ряду существенных проблем, которые можно решить с помощью автоматизации:

* Низкая скорость выполнения операций;
* высокая вероятность ошибки при запуске программ на множестве фалов: часть файлов может быть пропущена, а часть проанализирована более одного раза;
* очень низкая масштабируемость: при росте количества обрабатываемых файлов, которых в филогенетических исследованиях могут быть сотни, описанные выше проблемы кратно усугубятся.

Для более наглядной иллюстрации проблемы бизнес-процессов в текущем состоянии, рассмотрим детализацию одного из самых сложных процессов.

1. Детализация структурных диаграмм

Diagram

Description automatically generated Для детализации процесса А2 была использована нотация EPC.

Рисунок 5 – А2: Выделение и обработка кластеров ортологичных генов As-Is

Данная диаграмма демонстрирует, что в текущих процессах роль «Исследователь» перегружена, поскольку на неё опирается почти каждая операция, что неэффективно. Более того, по диаграмме видно, с каким большим количеством файлов приходится работать в рамках одного процесса, откуда следует, что при большом количестве входных данных, как и ожидается, можно ожидать не только сильного замедления работы, но и экспоненциального роста числа ошибок.

Diagram

Description automatically generatedРассмотрим, как преобразится этот процесс при автоматизации на рисунке 6.

Рисунок 6 - А2: Выделение и обработка кластеров ортологичных генов To-Be

Автоматизация позволила распределить операции исследователя между двумя компонентами системы, что должно не только значительно ускорить работу систем, но и исключить вероятность ошибок.

На основе результатов моделирования процессов сформулируем требования к разрабатываемой системе, в которой все операции исследователя автоматизированы.

1. Требования к ИС

Цель разрабатываемой системы – автоматизация филогенетического исследования геномов. Взаимодействие пользователя с системой будет ограничено вводом входных данных и получением отчётов с результатами. Все промежуточные операции будут делаться автоматически.

* Пользователи ИС: исследователи (биоинформатики).
* Потенциальный заказчик: государственный исследовательский институт.
* Интерфейс: консольное приложение.
* Платформа: ПК или вычислительный кластер на ОС из семейства Unix.
* Язык программирования: Python.

Для достижения цели система должна отвечать следующим функциональным требованиям:

1. Ввод и проверка входных данных: обязательно геномов и опционально геномных аннотаций.
2. Аннотация геномов.
3. Выделение кластеров ортологичных групп белков (КОГ).
4. Выделение кодирующих последовательностей геномов.
5. Выравнивание ортологичных последовательностей.
6. Построение филогенетического дерева исследуемых геномов.
7. Поиск генов под отбором.
8. Генерация отчётов о филогенетике исследуемых организмов. Отчёты должны содержать: филогенетическое дерево исследуемых генов, таблицу со статистикой по обнаруженным генам под отбором.

Также система должна отвечать следующим нефункциональные требованиям:

1. Удобство использования:
   1. Понятный и удобный пользователю интерфейс.
2. Надежность:
   1. Устойчивость к отказам.
3. Удобство сопровождения:
   1. Удобство внесения изменений.
   2. Анализируемость.
   3. Стабильность.

Diagram

Description automatically generatedНа основе анализа функциональных требований была составлена диаграмм вариантов использования системы, изображённая на рисунке 7.

Рисунок 7 - Диаграмма прецедентов

1. Требования к данным
   1. Требования к входным данным

На вход системе должно обязательно подаваться как минимум три непустых файла с геномами в формате FASTA. Пример файла в этом формате изображён на рисунке 8. В нём название последовательности следует за символом «>», после чего на следующей строке начинается текст статьи.

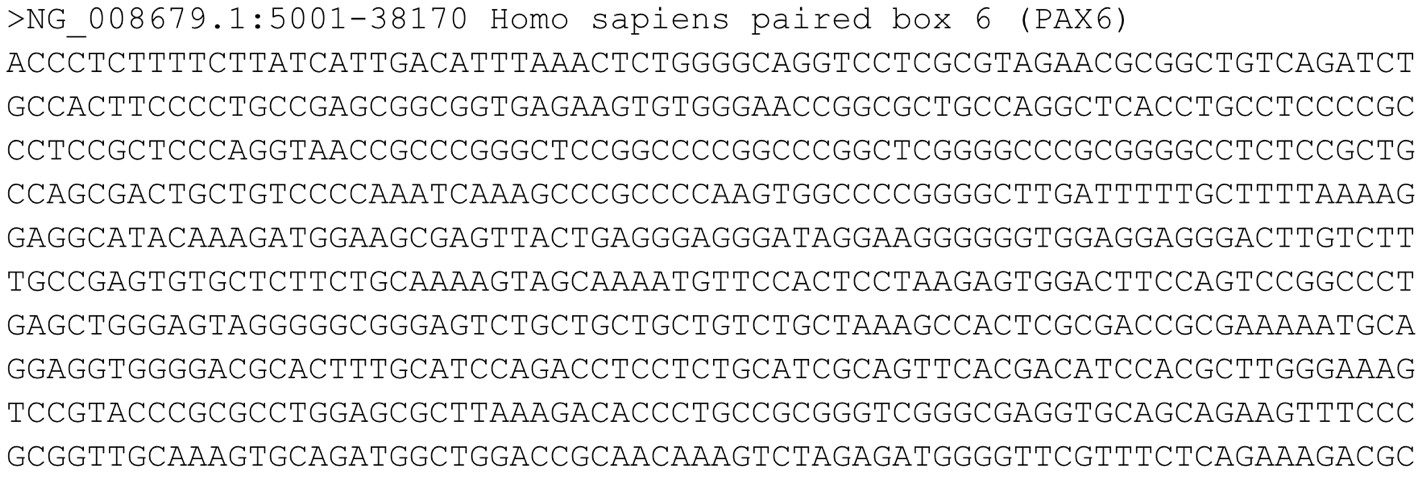


Рисунок 8 - Пример FASTA файла

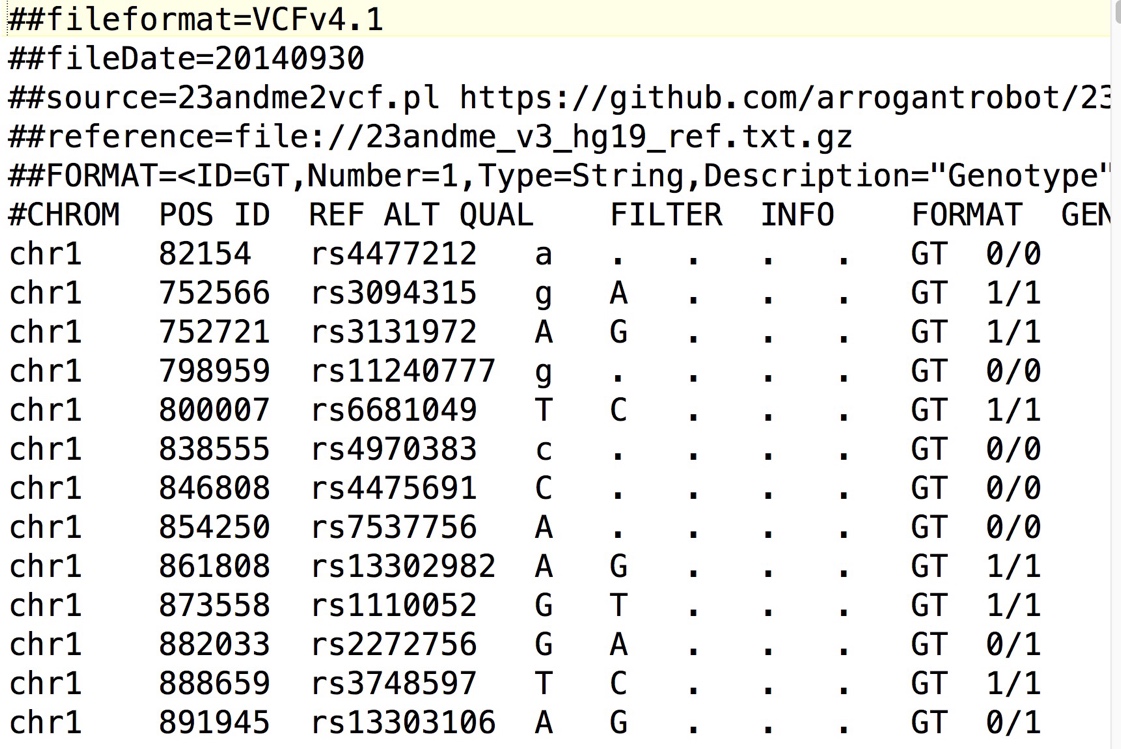
Опционально можно подать на вход файлы с аннотациями геномов в формате BED, пример которого представлен на рисунке 9.

Рисунок 9 - пример BED файла

* 1. Требования к выходным данным

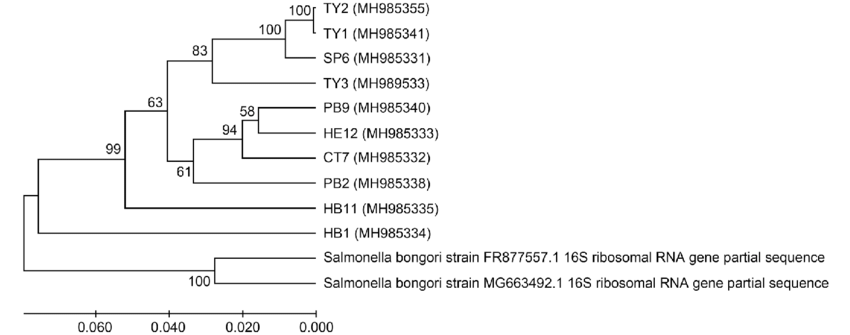
Во-первых, на выход программа должна выдавать файл с изображением филогенетического дерева геномов в формате PNG, подобное представленному на рисунке 10. На ветвях располагаются названия видов, заданные в файлах геномов.

Рисунок 10 - Пример отчёта с деревом

Table

Description automatically generatedВо-вторых, программа должна выводить таблицу в формате CSV, представляющую матрицу, в которой и по строкам, и по столбцам идут названия генов, а в ячейках расположены значения соотношений не синонимичных и синонимичных замен для рассматриваемой пары генов. Пример такой матрицы представлен на рисунке 11.

Рисунок 11 - пример отчёта по отбору генов

Глава 2. Проектирование

1. Проектирование архитектуры ИС
   1. Диаграмма компонент

Graphical user interface, application

Description automatically generatedНа основе анализа процесса филогенетического исследования геномов, описанного в первой главе, составим диаграмму компонентов ИС, показанную на рисунке 12.

Рисунок 12 - Диаграмма компонентов

На диаграмме можно выделить четыре типа компонент: специальные биоинформатические программы с расширением exe, скрипты на Python для промежуточной обработки данных с расширением py, входные данные (геномы) с расширением fna и выходные данные с расширениями png и csv.

С точки зрения связей между компонентами все достаточно типично для архитектуры конвейера: первый компонент зависит от входных данных, последующие компоненты зависят от предыдущих, а результат зависит от последнего вычисляющего компонента. Однако в представленной диаграмме есть и нетипичный для конвейера момент: скрипт «Cluster\_extractor.py» зависит от трёх компонент, поскольку должен объединять результаты первых двух этапов анализа.

В ходе проектирования и описания архитектуры ИС, было обнаружено, что использование стандартной нотации диаграммы компонент для описания биоинформатического вычислительного конвейера не оптимально, поскольку диаграмма получается не наглядной: без прочтения названий каждой компоненты непонятно, какие из них – специализированные вычислительные инструменты, а какие – программы для промежуточной обработки данных. Более того, прочтения названия может быть недостаточно, чтобы разобраться, где специальные программы, ведь читающий может не знать всех используемых в конвейере программ. Решить описанную проблему можно с помощью расширения нотации диаграммы компонент.

* 1. Диаграмма компонент в расширенной нотации

Diagram

Description automatically generatedПредлагаемое решение проблемы не наглядности стандартной нотации– добавить новый элемент в язык нотации, а именно компонент «биоинформатическое ПО». Данный компонент имеет все те же свойства, что и обычный компонент, но обозначается символом, наглядно сообщающим о том, что этот компонент – стороннее по отношению к системе биоинформатическое ПО. Диаграмма компонентов в расширенной нотации представлена на рисунке 13.

Рисунок 13 - Расширенная диаграмма компонентов

Нетрудно заметить, что диаграмма стала намного нагляднее и легко воспринимается без прочтения названий каждой компоненты.

1. Проектирование процессов ИС

Для проектирования процессов внутри разрабатываемой системы использовались диаграммы активностей, которые позволяют наглядно иллюстрировать взаимодействие компонент системы, а также описывать ветвления и параллельные процессы. Ниже представлены диаграммы активности всех процессов второго уровня детализации из структурной модели (процессы А1 – А4).

* 1. Graphical user interface, text

     Description automatically generatedПроцесс подготовки входных данных
  2. Graphical user interface

     Description automatically generatedПроцесс выделения и обработки кластеров ортологичных генов
  3. Graphical user interface, text, application, chat or text message

     Description automatically generatedПроцесс подготовки данных для отчётов
  4. Graphical user interface, text, application, chat or text message

     Description automatically generatedПроцесс создания отчётов

На основе спроектированных компонентов и процессов разработаем систему.

Глава 3. Разработка и тестирование

1. Принципы разработки конвейеров для анализа данных

A picture containing text, white goods, different, several

Description automatically generatedЛюбой вычислительный конвейер для анализа данных представляет собой направленный ациклический граф, структура которого описывает в каком порядке файлы с промежуточными результатами обрабатываются программами. Пример такого графа в общем виде изображён на рисунке 14[8]. В этом графе файлы являются узлами, а программы – переходами.

Рисунок 14 - Пример графа анализа

Таким образом, разработка конвейера для анализа данных сводится к описанию графа анализа, задающего правила, по которым программы передают файлы между друг другом.

Существуют два основных способа описания графа анализа, которые отличаются по способу описания структуры графа: явный (эксплицитный) и неявный (имплицитный). В эксплицитном способе описание структуры строится вокруг используемых программ: явным образом каждой программе присваиваются источники входных данных (другие программы). В имплицитном описании способе описание структуры строится вокруг данных: описывается последовательность результатов работы программ, а переходы в графе строятся автоматически исходя из описанных правил получения результатов.

1. Разработка конвейера для анализа данных

Для разработки вычислительных конвейеров существуют десятки инструментов, большая часть которых, к сожалению, не малоизвестны и не поддерживаются разработчиками. Инструмент для создания воспроизводимых и масштабируемых конвейеров для анализа данных Snakemake выгодно отличается от своих конкурентов тем, что имеет большую базу пользователей, подробную документацию и регулярные обновления. Кроме того, Snakemake очень часто используют в разработке биоинформатических программ, а в его основе лежит синтаксис, подобный синтаксису Python, что значительно упрощает освоение инструмента [8].

В Snakemake используется имплицитный способ описания структуры графа анализа, что не удивительно, поскольку данные – ключевая компонента любого исследования в области биоинформатики. На основе спроектированной во второй главе архитектуры ИС была описана структура графа анализа разрабатываемого конвейера на языке Snakemake (см. приложение А).

Компоненты, являющиеся сторонними программами, были загружены с официальных сайтов или репозитариев разработчиков. Оставшейся компоненты были написаны на языке Python с использованием библиотек из Biopython, проекта, объединяющего некоммерческие программные инструменты с открытым исходным кодом для вычислительной биологии. Исходный код всех компонент и представлен в приложении А.

1. Тестирование
   1. Запуск программы

Было проведено системное тестирование приложения на соответствие функциональным требованиям по критериям чёрного ящика: на вход были поданы геномы 7 видов вирусов, образующих семейство морбилливирусов. Каждый из видов специализирован заражать определённый вид животных: человека, у которого вызывает корь, собаку, кота, тюленя и т. д. Интересно посмотреть, в каких эволюционных отношениях находятся эти родственные, но разные виды вирусов и есть ли в них гены под отбором.

Text

Description automatically generatedПеред запуском программы убедимся, что папки с результатами «output» нет в рабочей директории перед началом тестирования. Затем запустим разработанную программу, хранящуюся в файле «Snakemake», передав ей папку с геномами вирусов в качестве входных данных, как представлено на рисунке 15.

Рисунок 15 - Запуск и работа программы

Text

Description automatically generatedПосле завершения обработки всех 7 геномов, которая не поместилась на скриншот, в рабочей директория появилась папка с результатами, в которой расположились непустые файлы с отчётами.

Рисунок 16 - Папка с результатами

* 1. Отчёт с филогенетическим деревом

Graphical user interface, text, application, email

Description automatically generatedФилогенетическое дерево всех видов семейства морбилливирусов изображено на рисунке 17.

Рисунок 17 - Отчёт с филогенетическим деревом

Дерево показывает степень родства между видами и их эволюционную историю. На основе этого дерева можно сделать много выводов о происхождении отдельных видов. Например, верхние три ветви дерева свидетельствуют о том, что вирус кори, заражающий человека, мог произойти от вируса кори козла, а тот в свою очередь от вируса, поражающего коров.

* 1. Гены под отбором

Graphical user interface, application, table, Excel

Description automatically generatedНа статистически значимом уровне в соответствии с тестом «Codon Based Z-test» под отбором оказались гены, представленные на изображении второго отчёта. В таблице и в столбцах, и строках располагаются гены, которые сравниваются друг с другом попарно [6].

Рисунок 18 - Отчёт с генами под отбором

Соотношения не синонимичных замен и синонимичных выделены синим цветом. Из соотношений видно, что заметная часть генов, чьё соотношение по модулю близко к единице, находятся под нейтральным отбором при попарном сравнении видов. Однако есть и гены с коэффициентами по модулю меньше единицы - они должны быть под отчищающим отбором, исключающим вредные для организма мутации. Среди нескольких генов под положительным отбором особенно выделяется ген с коэффициентом, равным 7,06 по модулю. Судя по всему, этот ген находится под сильным положительным отбором, должно быть его обладатель активно адаптируется к новым условиям среды.

Таким образом, тестирование прошло успешно – программа отвечает своим функциональным требованиям.

Заключение

В заключение работы хотелось бы отметить, что поставленная цель была достигнута – была разработана информационная система для автоматизации филогенетического исследования геномов. Каждая из поставленных задач была выполнена. Разработанная система удовлетворяет всем функциональным и не функциональным требованиям и позволяет автоматически проводить полное филогенетическое исследование, лишь введя исходные данные.

При изучении теоретического материала об этапах современного филогенетического исследования и используемых для этого программах наиболее полезным оказалась статья "Aquatic adaptation and depleted diversity: a deep dive into the genomes of the sea otter and giant otter" [4], в которой детально описано филогенетическое исследование геномов морских выдр. Другим важным источником была статья "Statistical methods for detecting molecular adaptation", описывающая статистические методы, используемые для выявления генетических элементов, подвергающихся естественному отбору [6].

Несмотря на значительный объём проделанной работы, разработанная система сохраняет пространство для улучшения и дополнения в разных аспектах. Например, текущий пользовательский интерфейс носит сугубо функциональный характер – можно сделать графический интерфейс, в том числе веб-интерфейс, который будет удобнее для пользователей, не привыкших работать с командной строкой. Кроме того, конвейер можно расширить, добавив в него больше программ для решения большего количество задач, например программу для автоматической оценки качества входных данных.

Библиографический список

1. Heywood, Vernon Hilton, and John McNeill, eds. Phenetic and phylogenetic classification: a symposium. No. 6. Systematics association, 1964.
2. Navarro, Fábio CP, et al. "Genomics and data science: an application within an umbrella." Genome biology 20.1 (2019): 1-11.
3. Lesk, Arthur M. Introduction to genomics. Oxford University Press, 2017.
4. Beichman, Annabel C., et al. "Aquatic adaptation and depleted diversity: a deep dive into the genomes of the sea otter and giant otter." Molecular biology and evolution 36.12 (2019): 2631-2655.
5. Tatusov, Roman L., et al. "The COG database: a tool for genome-scale analysis of protein functions and evolution." Nucleic acids research 28.1 (2000): 33-36.
6. Yang, Ziheng, and Joseph P. Bielawski. "Statistical methods for detecting molecular adaptation." Trends in ecology & evolution 15.12 (2000): 496-503.
7. Gollery, Martin. "Bioinformatics: sequence and genome analysis." Clinical Chemistry 51.11 (2005): 2219–2220.
8. Stepik [Электронный ресурс]: онлайн курс по управлению вычислениями от Института Биоинформатики, 2019. — Режим доступа: <https://stepik.org/course/1612/syllabus>, свободный. — Загл. с экрана.

ПРИЛОЖЕНИЕ A

***Исходный код приложения***

Исходный код всех компонентов приложения доступен на GitHub по ссылке: <https://github.com/Unknown-Negotiator/Pompan>.