Министерство науки и высшего образования РФ
Федеральное государственное автономное
образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт фундаментальной биологии и биотехнологии

Кафедра геномики и биоинформатики

УТВЕРЖДАЮ

Заведующий кафедрой

_____И.Е. Ямских

« ___» _____ 2024 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Два метода сравнения нуклеотидных последовательностей – выравнивания (alignment) и метода Шайдурова, проблемы и перспективы

06.04.01 – Биология

06.04.01.06 – Геномика и биоинформатика

Руководитель _____ д.ф.-м.н., проф. М.Г. Садовский
подпись, дата должность, ученая степень инициалы, фамилия

Выпускник _____ А.А. Тетерлева
подпись, дата инициалы, фамилия

Рецензент _____ д.ф.-м.н. С.И. Барцев
подпись, дата ученая степень инициалы, фамилия

Красноярск 2024

СОДЕРЖАНИЕ

РЕФЕРАТ

Выпускная квалификационная работа по теме «Два метода сравнения нуклеотидных последовательностей – выравнивания (alignment) и метода Шайдурова, проблемы и перспективы» содержит 55 страниц текстового документа, 10 иллюстраций, 2 таблицы, 86 источников.

Ключевые слова: СРАВНЕНИЕ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ВЫРАВНИВАНИЕ, НУКЛЕОТИДНЫЕ ПОСЛЕДОВАТЕЛЬНОСТИ, СВЁРТКА, МЕТОДЫ БЕЗ ВЫРАВНИВАНИЯ, МЕТОД ШАЙДУРОВА

Цель исследования — проанализировать два метода, используемые в задачах сравнения нуклеотидных последовательностей: выравнивание (alignment) и метод свёрточных функций (метод Шайдурова).

Объектом исследования в данной работе являются подходы к сравнению символьных последовательностей. Предметом исследования являются характеристики методов сравнения нуклеотидных последовательностей.

В результате проделанной работы было отобрано 25 митохондриальных геномов рыб из общедоступной базы данных NCBI и проанализированы филогенетические отношения между ними методом выравнивания и построения филогении на его основе, а также методом свёрточных функций (методом Шайдурова). Топологии деревьев, полученных методом максимального правдоподобия и методом присоединения ближайшего соседа были практически идентичны между собой, однако не показали абсолютное совпадение с топологией дерева, построенного на основе значений свёртки. Помимо того, была определена линейная зависимость (корреляция) между результатами (id и sim) попарного сравнения оригинальной последовательности гена TRAF6 человека с мутантными последовательностями данного гена и аналогичного сравнения, реализованного методом Шайдурова. Результаты показали, что между показателями свёртки и выравнивания при поиске мутаций точечного

типа, типа трактов (точковые мутации подряд) и инвертированных трактов наблюдается высокая степень корреляции.

ВВЕДЕНИЕ

В организме ДНК является генетическим материалом, который является носителем генетической информации, передаваемой от одного поколения к другому. Все живые существа со временем отличаются от общего предка в силу мутаций, обеспечивающих изменения в их ДНК. Возможность секвенирования ДНК организма является одним из наиболее важных и основных требований в биологических исследованиях. Раньше исследователи секвенировали несколько десятков или сотен нуклеотидов за раз, однако в настоящее время секвенирование ДНК выполняется высокопроизводительными машинами секвенирования, обеспечивающими сиквенс до 10^9 нуклеотидов в день. С появлением методов секвенирования «следующего поколения» (NGS), пропускная способность производства последовательностей увеличилась во много раз, а затраты на секвенирование снизились на порядки. Сравнение генома и метагенома, основанное на больших объемах данных секвенирования.

В настоящее время вычислительная биология и биоинформатика являются очень интересными и быстро развивающимися областями, где используется большое количество методов анализа отсеквенированных последовательностей. Одним из самых распространенных методов сравнения двух и более последовательностей, выявляющих степень схожести между ними, является выравнивание (alignment). Данный метод был предложен в 1970 году Нидельманом-Вуншем для аминокислотных и нуклеотидных последовательностей. С момента создания до настоящего времени выравнивание последовательностей используется практически повсеместно, особенно для изучения филогенетических отношений биологических последовательностей. Также следует отметить развитие данного метода с развитием технического прогресса, в том числе развитие алгоритмов множественного выравнивания,

однако все еще есть такие моменты работы алгоритма, которые ставятся под сомнение некоторыми исследователями, в том числе, например, наличие свободных параметров. Для решения данных проблем в последнее время большое развитие получают различные методы, не использующие выравнивание (AF-методы).

В данной работе приводятся особенности, перспективы и сравнение работы двух алгоритмов сравнения биологических последовательностей: метода выравнивания и относительно нового метода Шайдурова, который, как утверждается, может альтернативно использоваться для решения схожих задач в биоинформатике.

Целью данной работы является сравнение двух методов, используемых в задачах сравнения нуклеотидных последовательностей: выравнивания (alignment) и метода свёрточных функций (метод Шайдурова). Для достижения данной цели были поставлены следующие задачи:

1. Подобрать подходящий генетический материал (нуклеотидные последовательности), на котором на будет проводиться сравнение работы методов;
2. Сгенерировать набор изменённых нуклеотидных последовательностей, содержащих заданный набор мутаций, и провести сравнения исходной последовательности с изменёнными методом выравнивания и методом Шайдурова;
3. Получить результаты множественного выравнивания, провести филогенетический анализ результатов работы двух методов;
4. Провести сравнительный корреляционный анализ показателей результатов, получаемых двумя разными методами.

Глава 1. Обзор литературы

1.1. Сравнение последовательностей: имеющиеся подходы

Сравнительный анализ последовательностей ДНК и аминокислот имеет фундаментальное значение в биологических исследованиях, особенно в молекулярной биологии и геномике. Это первый и ключевой шаг в молекулярно-эволюционном анализе, предсказании функций генов и регуляторных областей, сборке последовательностей, поиске гомологии, предсказании молекулярной структуры, открытии генов и анализе взаимосвязей между структурой и функцией белка. Традиционно сравнение последовательностей основывалось на парном (PSA) или множественном (MSA) выравнивании последовательностей.

В традиционном подходе для сравнения последовательностей часто используются методы попарного и множественного выравнивания, такие как BLAST [1] и CLUSTAL [2]. Однако при работе с большими объемами данных, доступными благодаря новым технологиям секвенирования [3], эти методы могут оказаться неэффективными. Особенно сложно проводить анализ в случаях низкой идентичности последовательностей [4], например, при изучении отдаленно родственных белков или регуляторных областей генов [5,6], или для отдаленно родственных белковые гомологи [4] Также алгоритм выравнивания предполагает, что порядок сходства (гомологии) сохраняется внутри сравниваемых последовательностей линейно, поэтому выравнивания не учитывают возможные перестановки в последовательностях, такие как рекомбинации и замены белковых доменов [7] или горизонтальный перенос генов [8], и не применим случаях, когда обрабатываются крупномасштабные наборы данных о последовательностях, например, для филогенетики всего генома [9]. На практике невозможно точно сравнить две длинные последовательности ДНК длиной в миллионы нуклеотидов. Таким образом, в современных исследованиях возникает необходимость в разработке более точных и эффективных методов анализа биологических последовательностей.

Поэтому в качестве альтернативы выравниванию последовательностей было разработано множество так называемых подходов к анализу последовательностей без выравнивания (alignment free, AF) [4], причем самые ранние работы относятся к середине 1970-х годов [10], хотя концепция «alignment free» сравнения последовательностей привлекло повышенное внимание только в начале 2000-х годов [11]. Большинство этих методов основаны на статистике слов или сравнении слов, а их масштабируемость позволяет применять их к гораздо большим наборам данных, чем традиционные методы на основе выравнивания.

1.1.1. Выравнивание (alignment)

Выравнивание (alignment) последовательностей обычно является первым шагом, выполняемым в биоинформатике для понимания молекулярной филогении неизвестной последовательности, для определения биологической функции молекулы, для выявления подобия в сравниваемых последовательностях, для поиска возможных предков. Это делается путём согласования неизвестной последовательности с одной или несколькими известными последовательностями, взятыми из базы данных для выделения общих частей, поскольку нуклеотиды, определяющие важную функциональную и структурную роль, как правило, сохраняются в ходе естественного отбора в процессе эволюции. Оптимальное выравнивание выстраивает две или более последовательностей таким образом, чтобы совпадало максимальное количество идентичных или подобных остатков. Последовательности могут быть нуклеотидными (ДНК или РНК) или аминокислотными (белки). Процедура выравнивания может привести к появлению одного или нескольких пробелов или зазоров (вставок или делеций) в сравниваемых последовательностях. Выявление делеции (на первом шаге определяемой как наличие одного или нескольких пробелов в одной из сравниваемых последовательностей) указывает на возможную потерю; таким образом, эволюционная вставка или делеции,

транслокации и инверсии наряду с заменами нуклеотидов в индивидуальных позициях (SNP) в последовательности составляют набор тех мутаций, которые подлежат выявлению в сравниваемых последовательностях.

Выравнивание или картирование является важным шагом после процедуры секвенирования. При этом возможны два типа выравнивания, а именно: попарное выравнивание последовательностей (PSA) и выравнивание нескольких последовательностей (MSA). PSA сравнивает только две последовательности одновременно, тогда как MSA выравнивает несколько (более двух); как правило, MSA используется для определения родственных связей в семействе последовательностей. MSA также является предпосылкой сравнительного геномного анализа для идентификации и количественной оценки консервативных областей или функциональных мотивов в целом семействе последовательностей, оценки эволюционного расхождения между последовательностями. Выравнивание последовательностей на аминокислотном уровне может иметь в некоторых случаях большее значение, чем сравнение на нуклеотидном уровне, поскольку белок является ключевой функциональной биологической молекулой, несёт структурную и/или функциональную информацию, постольку можно ожидать меньшей изменчивости в аминокислотных последовательностях. Выравнивание, таким образом, также имеет тесную связь со структурной биологией. Следовательно, выравнивание, в частности MSA, является отправной точкой во многих биологических макромолекулярных исследований.

При выравнивании последовательности используются различные методы оценки, чтобы узнать уровень идентичности или сходства. Нуклеотидная оценка — это простая схема идентификации, при которой идентичным основаниям в обеих последовательностях присваиваются положительные баллы. Напротив, для белка также подсчитывается оценка сходства (наряду с оценкой идентичности), обозначающая аминокислоты, имеющие сходные физико-химические свойства. Матрицы замещения, к которым чаще всего обращаются

для выравнивания белковых последовательностей, это точечная принятая мутация (PAM) и матрица BLOcked SUbstitution Matrix (BLOSUM).

Метод (alignment) может быть двух видов: глобальное выравнивание [12] и локальное выравнивание [13]. Глобальное выравнивание выполняется, когда сходство подсчитывается по всей длине последовательностей. Некоторые методы MSA достигают глобального выравнивания, но трудности возникают, когда последовательности гомологичны только в локальных областях, где четкий блок неиспользованного выравнивания является общим для всех последовательностей, или если есть присутствие перетасованных доменов среди связанных последовательностей. В таких случаях выполняется локальное выравнивание, чтобы узнать локальные сходные области среди последовательностей. Когда существует большая разница в длинах сравниваемых последовательностей, обычно выполняется локальное выравнивание.

Для решения задач парного выравнивания длинных последовательностей разумно использовать возможности динамического программирования. Сущность задачи заключается в том, что компьютер проводит оценку всех возможных выравниваний, связывая последовательности между собой во всех вариантах. На основе этой оценки, выраженной в форме «баллов» (score), присуждаемых или вычитаемых за различные варианты, предлагается оптимальное соотнесение представленных строк. Исходная задача может показаться простой, однако в случае ДНК(РНК) и белков могут возникать разнообразные замены, повторы и инделы (пропуски) в последовательностях, которые могут быть недостаточно отчетливо представлены в сиквенсах по различным причинам, порождая множество различных вариантов соотнесения двух строк. Даже при использовании мощных компьютеров такие выравнивания могут занимать продолжительное время.

Несмотря на высокое качество сравнения при использовании алгоритмов, основанных на алгоритмах локального и глобального выравнивания, затраченное время на вычисления зачастую не пропорционально полученным

данным. В таких случаях приходят на помощь эвристические и вероятностные методы обработки данных. Эти методы, хотя и не обеспечивают высокую точность в сравнении с вышеописанными методами, значительно экономят вычислительные ресурсы.

Эвристические алгоритмы основаны на функции, которая упорядочивает альтернативы на каждом этапе разветвления вариантов (упорядочивание происходит на основе изначально установленных и введенных данных, то есть, основываясь на имеющейся информации), что позволяет получить приблизительно оптимальный ответ. Такой подход не является абсолютно точным, но остается ценным благодаря скорости получения результата. Вся эвристика оперирует на подобных принципах.

Вероятностные же методы, помимо использования основных алгоритмов выравнивания учитывают разные вероятности тех или иных замен. В этом случае строятся так называемые матрицы замен. Разные аминокислоты заменяются в процессе эволюции с разной вероятностью. И для учёта этой неравной вероятности замен используются матрицы. Не вдаваясь в частности, отметим, что и у этих методов есть свои недостатки. Так, например, точность вероятностных методов сильно уменьшается с увеличением дистанции между таксонами. Кроме эвристик, для решения ряда проблем могут быть использованы различные AF-методы.

Кроме того, проблемы для подходов, базирующихся на выравнивании, может вызвать сравнение генома и метагенома, основанное на больших объемах данных секвенирования нового поколения (NGS), из-за огромного объема данных и относительно короткой длины считывания. Подходы без выравнивания в таком случае, как правило, эффективны с точки зрения вычислений. Эти подходы применены ко многим современным проблемам биоинформатики, включая сравнение регуляторных областей генов, последовательностей генома, метагеномов, биннинг контигов в метагеномных данных.

1.1.2. Проблемы выравнивания

Выравнивание биологических последовательностей является одним из ключевых инструментов в анализе генетической информации. Вне зависимости от используемого метода, программы для выравнивания в своей основе ищут сопоставления между отдельными нуклеотидами или аминокислотными остатками, расположенными в одном и том же порядке в двух или более последовательностях. Этот процесс предполагает классификацию каждого символа последовательности как идентичного/похожего (совпадение) или неконсервативного (несовпадение), хотя большинство программ также учитывают вставленные/удаленные состояния (инделы, гэпы). Однако, с расширением наших знаний о биологических последовательностях и их эволюционных особенностях, стали видны определенные недостатки в использовании только выравниваний для сравнения последовательностей. Все эти проблемы усугубляются в полногеномном масштабе, и лишь немногие из них можно решить за счёт увеличения вычислительной мощности или более совершенных моделей замен.

Ниже приведены ситуации, когда анализ последовательностей на основе выравнивания может вызвать определенные проблемы:

Проблема 1. Программы, производящие выравнивание, предполагают, что гомологичные последовательности содержат ряд линейно расположенных и более или менее консервативных участков последовательностей. Однако это предположение, которое называется коллинеарностью, очень часто нарушается в реальном мире. Хорошим примером являются вирусные геномы, которые демонстрируют большие различия в количестве и порядке генетических элементов из-за их высокой частоты мутаций, частых событий генетической рекомбинации, горизонтальных переносов генов, дупликации генов и приобретений/потерь генов [14]. Эти крупномасштабные эволюционные процессы, по существу, постоянно происходят в геномах других организмов.

Проблема 2. При анализе биологических последовательностей с использованием выравнивания возникают сложности, особенно когда идентичность между последовательностями опускается ниже определенного уровня. Для белковых последовательностей, состоящих из 20 различных аминокислот, даже две случайные последовательности могут совпадать лишь на 5% остатков. При учете вставок и удалений этот процент может увеличиться до 25% [15]. Таким образом, диапазон идентичности от 20% до 35% обычно считается «сумеречной зоной» [16], где трудно различить дальних родственников от случайных последовательностей. Если сходство между последовательностями опускается ниже 20% в этой «сумеречной зоне», то задачи поиска гомологии становятся сложными для определения с помощью стандартных методов попарного выравнивания. В таких случаях требуются более сложные подходы, такие как использование профилей (например, PSI-BLAST) или скрытых марковских моделей (например, HMMER). Эта проблема особенно актуальна при аннотации белковых суперсемейств, где члены могут сохранять структурное сходство даже при низкой идентичности в 8–10% [17]. Для нуклеотидных последовательностей точность выравнивания представляет еще большие требования. Например, две случайные ДНК/РНК последовательности могут иметь до 50% идентичности, если учитывать вставки и удаления. Граница области «сумеречной зоны» может охватывать совпадения в нуклеотидах до 60–65% [18].

Проблема 3. Подходы, основанные на выравнивании, как правило, занимают много памяти и требуют больших вычислительных ресурсов, что ограничивает применение данных методов для сравнения большого числа геномов. Число возможных вариантов выравнивания двух последовательностей быстро растет с длиной последовательностей. Хотя существует способы оптимизации процесса, например, при использовании динамического программирования, что гарантирует получение математически оптимального выравнивания с наибольшими баллами, не прибегая к перечислению всех возможных решений. Данный метод также требует больших вычислительных

ресурсов (временная сложность имеет порядок произведения длин входных последовательностей) [19]. Поэтому, несмотря на богатство инструментария и на опыт многолетних исследований [20], проблема выравнивания длинных последовательностей до конца не решена [21]. Кроме того, доступные эволюционные модели последовательностей могут не применяться непосредственно к полным геномам.

Проблема 4. Задача вычисления точного выравнивания нескольких последовательностей является NP-трудной. Это значит, что такое выравнивание невозможно выполнить в разумные сроки из-за необходимости рассмотрения всех возможных вариантов. Как следствие, было разработано более ста более быстрых альтернативных методов за последние десятилетия [22]. Однако, скорость имеет свою цену. Эти методы основаны на различного рода эвристиках и нацелены на нахождение в целом оптимального, но не гарантированно лучшего и точного выравнивания с максимальным количеством баллов, что иногда приводит к неточностям и снижению качества работы. Последнее в итоге может ограничивать качество многих последующих анализов, например, филогенетического.

Проблема 5. Выравнивание последовательностей зависит от множества априорных предположений об эволюции сравниваемых последовательностей. Эти различные свободные параметры (к примеру, матрицы замещения, штрафы за внесение или продолжение разрыва в выравнивание, а также пороговые значения для статистических параметров) в каком-то смысле произвольны. Более того, система подсчета баллов не согласована между различными инструментами и приложениями, и многие отчеты показали, что даже небольшие изменения во входных параметрах потенциально могут сильно повлиять на получаемое в итоге выравнивание [23]. Несмотря на понимание существования данных проблем, выбор параметров для получения выравнивания может часто вызывать проблемы у исследователя и, как правило, требует метода проб и ошибок (т.е. если выравнивание получилось недостаточно хорошее, то можно настроить входные параметры так, чтобы получить более «привлекательные»

результаты). Кроме того, матрицы замещения, которые необходимы для выравнивания белков (например, различные матрицы замещения BLOSUM и PAM), часто используются без проверки на репрезентативность для выравниваемых последовательностей.

Глобальные и локальные алгоритмы выравнивания могут быть работать некорректно с несоответствиями, пробелами, чередующимися блоками в биологических последовательностях, инверсиями, которые легко найти в практически любом генетическом материале. Эти методы могут сделать ошибочный вывод о том, что функционально связанные последовательности на самом деле в значительной степени не связаны между собой, поскольку они не демонстрируют какого-либо статистически значимого выравнивания. Длина последовательности также важна при выполнении выравнивания из динамического программирования. Например, локальные и глобальные, реализованные в программах типа ClustalW [24], имеют сложность, сильно возрастающую при увеличении длин входящих последовательностей, и поэтому понятно, что их требования к ресурсам быстро возрастают для больших последовательностей. Часто бывает нецелесообразно проводить сравнения полных геномов с помощью этого подхода из-за большого количества времени, которое это потребует времени. По этой причине приобрели популярность технологии, реализующие быстродействие в базах данных, такие как BLAST [25], BLASTZ [26] и BLAT [27]. Другие методы, помогающие преодолеть некоторые ограничения динамического программирования, пришли из различных областей, таких как облачные вычисления [28], распределенные вычисления [29] и параллельные вычисления для сравнения нескольких последовательностей [30].

За последние два десятилетия был разработан широкий спектр AF-подходов к сравнению последовательностей. Эти подходы включают методы, основанные на подсчете слов или k-меров [31], длине общих подстрок [32–34], микровыравниваниях [35–38], моментах положений нуклеотидов [39], преобразованиях Фурье [40], теории информации [41] и системах итерированных

функций [42,43]. В настоящее время наиболее широко используемые подходы AF основаны на подсчете k-меров [43]. Эти методы очень разнообразны и обеспечивают множество статистических измерений, которые реализованы в различных программных инструментах [4,44–46]. Многие k-мерные методы работают путем проецирования каждой из входных последовательностей в пространство признаков количества k-меров, где информация о последовательностях преобразуется в числовые значения (например, частоты k-меров), которые можно использовать для расчета расстояний между всеми возможными парами последовательностей в данном наборе данных. Углубленный обзор методов сравнения последовательностей без выравнивания был рассмотрен в ряде публикаций [4,41,47–49].

Алгоритмы, основанные на частоте, которые управляются статистикой употребления слов или чем-то подобным, становятся популярными в недавних исследованиях. Это связано с тем, что эти подходы, как правило, не имеют проблем вычислений, вызванными несоответствиями, пробелами и инверсиями в последовательностях, которые часто встречаются между последовательностями для сравнения [50]. Например, эти методы функционируют, оценивая информационное содержание между последовательностями, и поэтому чередование блоков ДНК между двумя последовательностями не будет вызывать проблемы. Эта форма выравнивания не зависит от того, где в последовательности находятся объекты, а только от факта, что последовательность содержит эти объекты. Методы, использующие частотный анализ, также не обладают высокой алгоритмической сложностью, поскольку они, как правило, линейны. Таким образом, они способны обрабатывать большие последовательности с меньшими ресурсами, чем алгоритмы динамического программирования, и не полагаются на поддержку баз данных, как это было бы с BLAST, BLASTZ или BLAT, например. Очевидно, что существует потребность в альтернативном подходе к сравнению последовательностей, выполняемых методами, не относящимися к динамическому программированию, и поэтому методы без выравнивания

становятся особенно привлекательными для исследований в области биоинформатики, где данные являются естественно динамичными.

1.1.3. Методы сравнения последовательностей без выравнивания

Подходы к сравнению последовательностей без выравнивания могут быть определены как любой метод количественной оценки сходства/несходства последовательностей, который не использует и не производит выравнивание (присвоение соответствия или несоответствия остатка остатку) на любом этапе применения алгоритма. С самого начала такое ограничение ставит подходы без выравнивания в выгодное положение — поскольку методы без выравнивания не полагаются на динамическое программирование, они менее затратны в вычислительном отношении, поскольку они, как правило, имеют линейную сложность, зависящую только от длины последовательности запросов [48] и иногда используются для полногеномного сравнения [31,37,51]. Методы без выравнивания также устойчивы к событиям транслокаций и рекомбинации и применимы, когда существование низких гомологий последовательностей не может быть надежно обработано выравниванием [52]. Наконец, в отличие от методов, основанных на выравнивании, они не зависят от предположений относительно эволюционных траекторий изменений последовательностей. Всё вышесказанное применимо ко всем методам без выравнивания. Подходы без выравнивания можно условно разделить на две группы [11,41]: методы, основанные на частотах подпоследовательностей определенной длины (методы, основанные на словах), и методы, оценивающие информативность между полноразмерными последовательностями (методы, основанные на теории информации). Существуют также методы, не попадающие в эту классификацию, в том числе основанные на длине совпадающих слов [53], самые длинные общие [54] или минимальные отсутствующие [55,56] слова между последовательностями, итерированных картах [42], а также на графическом представлении последовательностей ДНК, которые количественно фиксируют

суть базового состава и распределения последовательностей [57]. Все подходы, свободные от выравнивания, математически хорошо обоснованы в области линейной алгебры, теории информации и статистической механики и вычисляют попарные меры несходства или расстояния между последовательностями. Удобно, что большинство из этих мер могут быть непосредственно использованы в качестве входных данных в стандартном программном обеспечении для построения деревьев, таком как, например, MEGA [58].

Несмотря на значительный прогресс, достигнутый в области сравнения последовательностей AF-методами [4], разработчики и пользователи этих методов сталкиваются с рядом трудностей. Новые AF-методы обычно оцениваются их авторами, и результаты публикуются вместе с этими новыми методами. Таким образом, трудно сравнивать эффективность этих инструментов, поскольку они основаны на несогласованных стратегиях оценки, различных наборах данных и переменных критериях тестирования. Как следствие, оценка новых алгоритмов отдельными исследователями в настоящее время отнимает значительное количество времени и вычислительных ресурсов, что усугубляется непреднамеренной предвзятостью сравнения.

Проблема оценки методов без выравнивания с увеличением количества и качества последних требует все больше внимания со стороны исследователей. Так авторы одной из статьи [59] разработали веб-сервис AFproject, дающий оценку новым и уже существующим AF-методам сравнения нуклеотидных последовательностей. Авторы предлагают оценку методов, основанную на работе с несколькими типами биологических данных: генов, белков, регуляторных элементов или геномов. Более того, оценка производится при различных эволюционных сценариях, например, высокая мутабельность или горизонтальный перенос генов (HGT).

1.2. Митохондриальный геном

Для демонстрации работы двух рассматриваемых методов, в том числе для построения филогенетических деревьев, был предложен митохондриальный геном некоторых видов рыб подотряда *Labroidei*.

Митохондриальный геном (mtDNA, мтДНК) представляет собой небольшую кольцевую, хотя у некоторых организмов встречаются линейные формы, двуцепочечную молекулу ДНК, находящуюся в митохондриях клеток — органеллах клеток эукариот, ответственных за производство энергии. Митохондрии играют ключевую роль в процессе производства энергии в клетках путем окисления пищевых веществ. У митохондрий есть своя собственная ДНК, отличная от ядерной ДНК клетки, помимо этого в растительных клетках существует также хлоропластная ДНК. Митохондриальный геном имеет ряд уникальных особенностей, отличающих его от ядерного генома [60,61].

У большинства животных мтДНК имеет размер около 16-18 тысяч пар оснований. Митохондриальный геном кодирует 37 генов у большинства животных, включая 13 белков, 22 транспортных РНК (тРНК) и 2 рибосомальных РНК (рРНК). Белки, кодируемые мтДНК, в основном участвуют в процессах окислительного фосфорилирования и синтеза АТФ. Репликация и транскрипция мтДНК отличаются от аналогичных процессов в ядерном геноме и регулируются отдельными митохондриальными полимеразами и транскрипционными факторами [62]. МтДНК, как правило, демонстрирует высокую скорость мутаций, что делает её полезной для исследований филогении и популяционной генетики [63]. Кроме того, мтДНК наследуется по материнской линии, что упрощает изучение генетической истории и происхождения видов [64].

Митохондриальный геном рыб также представляет собой кольцевую молекулу ДНК, содержащую гены для синтеза белков, необходимых для работы митохондрий, и так же, как и у других животных, имеет ряд характерных особенностей. Размер митохондриального генома рыб варьируется, но в среднем составляет около 16,5 тысяч пар оснований, как и у других позвоночных.

Генетическое содержание мтДНК рыб аналогично и включает 13 белков, 22 тРНК и 2 рРНК [65]. У рыб могут наблюдаться некоторые вариации в структуре и порядке генов, что связано с их широким эволюционным разнообразием [66]. В отличие от многих других позвоночных, у некоторых видов рыб обнаружены дополнительные уникальные гены или отсутствующие традиционные [67]. Митохондриальные гены характеризуются более высокой скоростью мутаций по сравнению с ядерными, что, вероятно, связано с воздействием свободных радикалов и активных форм кислорода, образующихся в дыхательной цепи митохондрий [68], а также с менее эффективной системой репарации мтДНК по сравнению с аналогичной ядерной системой. Несмотря на сильный очищающий отбор, который необходим для функционирования митохондриальных генов, высокая частота мутаций приводит как к соматическим изменениям, накапливающимся с возрастом, так и к мутациям зародышевой линии, вызывающим изменчивость внутри вида и дивергенцию между видами [69]. В последние десятилетия, благодаря таким характеристикам мтДНК, как высокая частота мутаций и практически полное отсутствие генетической рекомбинации, она стала одним из наиболее популярных маркеров для оценки генетического разнообразия видов [70]. Интересно, что митохондриальный геном рыб обычно менее подвержен мутациям по сравнению с митохондриальными геномами других организмов, что делает его ценным инструментом для исследований в области эволюции и генетики рыб. МтДНК также может использоваться для определения родства и идентификации видов рыб, а также для понимания адаптивных изменений и эволюционных процессов, происходящие в различных водных экосистемах [71,72].

1.3. Ген TRAF6 человека

Ген TRAF6 (TNF receptor associated factor 6) кодирует белок, который является ключевым участником сигнальных путей, связанных с иммунным ответом, воспалением и костной резорбцией [73]. Этот ген играет важную роль

в передаче сигналов от рецепторов опухолевого некроза (TNF) и других рецепторов, активируемых интерлейкинами. Ген TRAF6 находится на 11-й хромосоме человека (локус 11p12). Он состоит из нескольких экзонов, которые кодируют белок длиной 522 аминокислоты. TRAF6 содержит несколько важных доменов, включая RING-домен, который обладает Е3убиквитинлигазной активностью [74], и TRAF-домен, который участвует в белок-белковых взаимодействиях.

Основные функции TRAF6:

Сигнальная трансдукция: TRAF6 является адаптерным белком, который связывает рецепторы клеточной поверхности с внутриклеточными сигнальными каскадами. Он участвует в активации ядерного фактора каппа B (NF-κB) [74,75] и MAPK (митоген-активируемая протеинкиназа), которые регулируют экспрессию генов, связанных с воспалением и иммунным ответом.

Иммунный ответ: TRAF6 необходим для нормального функционирования врожденного и адаптивного иммунного ответа. Он участвует в передаче сигналов от Toll-подобных рецепторов (TLR) и рецепторов интерлейкина-1 (IL-1), которые играют ключевую роль в иммунной защите [76].

Костная резорбция: TRAF6 участвует в дифференцировке остеокластов — клеток, ответственных за разрушение костной ткани [77]. Этот процесс важен для поддержания костного гомеостаза и ремоделирования костей.

Дисфункция или мутации в TRAF6 могут привести к нарушению иммунного ответа, что связано с развитием различных аутоиммунных заболеваний, таких как системная красная волчанка (СКВ) и ревматоидный артрит [78]. Мутации в TRAF6 также могут вызывать нарушение дифференцировки остеокластов, что приводит к остеопетрозу — заболеванию, характеризующемуся повышенной плотностью костей и нарушением их структуры. Нарушения в сигнальных путях, связанных с TRAF6, могут способствовать развитию некоторых видов рака [79], включая рак груди и лейкемию, за счет дисрегуляции клеточного роста и апоптоза [80].

Глава 2. Материалы и методы

2.1. Генетический материал

Материалом для данного исследования послужила последовательность гена TRAF6 человека, его кодирующая часть. Последовательность была взята из базы данных NCBI. ID гена TRAF6 (TNF receptor associated factor 6) у человека — 7189. Далее исходная последовательность длиной 1569 нуклеотидов (CDS последовательности) была преобразована в последовательности–копии со следующими типами мутаций:

тип 1. Точечные мутации замены (SNP), расположенные случайно по всей длине последовательности.

тип 2. Совокупность точечных мутаций, образующих непрерывные «тракты».

тип 3. Инверсированные «тракты».

тип 4. Инделы (gap) в исходной последовательности длиной от 1 до 19 нуклеотидов.

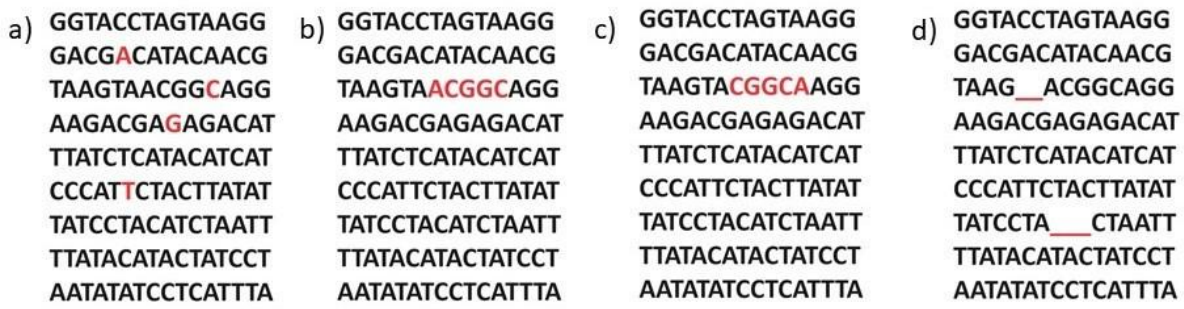

Рисунок 1. Примеры введённых мутаций. a)–пример точечных мутаций (SNPs); b)–пример тракта; c)–пример инвертированного тракта; d)–пример индела (gap).

Таблица 1 – Исследуемые митохондриальные геномы рыб. AC # — Accession number (GenBank); Species — видовое название; Classification — подотряд, семейство.

AC #	**Species**	**Classification**
NC_009063	*Tropheus duboisi*	*Labroidei, Cichlidae*
NC_009062	*Neolamprologus brichardi*	*Labroidei, Cichlidae*
NC_013750	*Oreochromis*	*Labroidei, Cichlidae*
NC_013663	*Oreochromis niloticus*	*Labroidei, Cichlidae*
NC_009057	*Oreochromis sp. KM_2006*	*Labroidei, Cichlidae*
NC_011171	*Tylochromis polylepis*	*Labroidei, Cichlidae*
NC_011168	*Hypselecara temporalis*	*Labroidei, Cichlidae*
NC_009058	*Astronotus ocellatus*	*Labroidei, Cichlidae*
NC_011169	*Ptychochromoides katria*	*Labroidei, Cichlidae*
NC_011170	*Paratilapia polleni*	*Labroidei, Cichlidae*
NC_011177	*Paretroplus maculatus*	*Labroidei, Cichlidae*
NC_011179	*Etroplus maculatus*	*Labroidei, Cichlidae*
NC_018814	*Petrochromis trewavasae*	*Labroidei, Cichlidae*
NC_018815	*Tropheus moorii*	*Labroidei, Cichlidae*
NC_009064	*Abudefduf vaigiensis*	*Labroidei, Pomacentridae*
NC_009065	*Amphiprion ocellaris*	*Labroidei, Pomacentridae*
NC_009059	*Cymatogaster aggregata*	*Labroidei, Embiotocidae*
NC_009060	*Ditrema temminckii*	*Labroidei, Embiotocidae*
NC_012055	*Pseudolabrus eoethinus*	*Labroidei, Labridae*
NC_009067	*Pseudolabrus sieboldi*	*Labroidei, Labridae*
NC_010205	*Pteragogus flagellifer*	*Labroidei, Labridae*
NC_009066	*Halichoeres melanurus*	*Labroidei, Labridae*
NC_009459	*Parajulis poecilepterus*	*Labroidei, Labridae*
***Osmeriformes* (outgroup)**		
NC_013564	*Alepocephalus agassizii*	*Alepocephaloidei, Alepocephalidae*
NC_013577	*Bajacalifornia megalops*	*Alepocephaloidei, Alepocephalidae*

Таким образом, общее число последовательностей для анализа включало 1 исходную последовательность без мутаций и 14 последовательностей для 4 типов мутаций: 4 последовательности с SNP, 4 последовательности с трактами, 4 последовательности с инвертированными трактами, 2 последовательности с инделами. Рисунок 1 иллюстрирует исследуемые в работе типы мутаций.

Для оценки работы методов также был использован генетический материал, включающий последовательности митохондриального генома рыб. Набор данных объединяет 25 собранных геномов видов рыб подотряда окунеобразных *Labroidei* (относится к числу наиболее богатых семействами и видами среди костистых рыб), взятых из Fisher et al., 2013 [81]. Геномные последовательности включают 25 видов рыб и демонстрируют типичную митохондриальную структуру с 13 генами, кодирующими белки, 2 генами рРНК, 22 генами тРНК и некодирующей контрольной областью. Перечисленные в таблице 1 полные митохондриальные геномы были загружены из базы данных нуклеотидов NCBI. Укоренение дерева проводилось на последовательности двух видов внешней группы: *Bajacalifornia megalops* и *Alepocephalus agassizii*.

2.2. Сравнение методом свёрточных функций (метод Шайдурова)

Альтернативой выравниванию, согласно ряду публикаций [82–84], может быть посимвольное сравнение последовательностей с подсчётом точных совпадений. При этом понятно, что число точных совпадений будет зависеть от того, какие фрагменты сравниваемых последовательностей рассматриваются. В идеале следует перебрать все возможные наложений фрагментов и для каждого подсчитать число точных совпадений. Анализ влияния несовпадений на результат сравнения будет рассмотрен ниже.

Очевидно, что прямой подсчёт числа совпадений (методом грубой силы) является чрезвычайно трудоёмкой с точки зрения вычислений задачей. Здесь вместо прямого подсчёта можно использовать различные методы, повышающие производительность вычислений.

Одним из высоко эффективных способов понизить трудоёмкость вычислений совпадений является подход, основанный на вычислении свёртки двух многочленов; для чего исходные символьные последовательности преобразуются в числовые (бинарные, {0,1}). Это преобразование делается следующим образом: все символы A заменяются на единицы, а все остальные символы — на нули. Затем берётся копия исходной символьной последовательности и в ней все символы C (и все остальные) заменяются по аналогичной процедуре. Тем самым, исходная символьная последовательность преобразуется в четыре бинарных. Этот приём показан на рисунке 2.

Будем рассматривать эти бинарные последовательности как коэффициенты многочлена степени N, где N — длина исходной последовательности. Тогда произведение двух бинарных последовательностей, в которых единицы являются заменами одинаковых символов, определяет {0,1} многочлен, коэффициенты которого соответствуют точным совпадениям однородных символов; заметим, что такое произведение является многочленом степени N1+N2, где N1 и N2 — длины сравниваемых последовательностей. Именно для вычисления такого произведения бинарных многочленов было предложено использовать их свёртку. Подробнее см. в разделе 2.2.1.

Свёрткой называют произведение двух полиномов, в котором один из сомножителей инвертирован (записан в противоположном порядке).

Воспользуемся бинаризацией исходных последовательностей; это приведёт к тому, что многочлен, являющийся произведением, также является бинарным (его коэффициентами являются единицы либо нули) и для его вычисления воспользуемся преобразованием Фурье сравниваемых бинаризованных последовательностей. На практике в этих вычислениях используется метод быстрого преобразования Фурье, что позволяет очень значительно понизить затраты на вычисление (как по времени, так и по вычислительным ресурсам). Подчеркнем, что применение быстрого преобразования Фурье позволяет за один прогон вычислить точное число совпадений во всех мыслимых наложениях последовательностей друг на друга.

Здесь следует также подчеркнуть, что получающаяся свёртка является числовой последовательностью, элементами которой являются значения точных совпадений однородных символов, определяемых в текущем наложении, вне зависимости от того, в каком именно месте этого наложения наблюдается точное совпадение.

2.2.1. Преобразование символьных последовательностей в числовые

Рассмотрим четырехбуквенный алфавит $\aleph$ = {A,C,G,T}. Пусть P = $\{p_k\}_{k=1}^{N}$ и Q= $\{q_k\}_{k=1}^{L}$ — две последовательности символов из $\aleph$ длины N и L соответственно, N $\geqslant$ L.

Заметим, что длины последовательностей и мощность алфавита могут быть произвольными. При реализации все ограничивается объёмами оперативной памяти, необходимой для проведения вычислений, и задачами исследования. Поскольку мы будем сравнивать генетические последовательности, то зафиксируем алфавит из четырёх символов.

Каждой из исходных последовательностей поставим в соответствие $|\aleph|$ бинарных последовательностей, полученных по принципу, показанному на Рис. 2 (для $|\aleph|$ = 4). Алгоритм формирования бинарных последовательностей следующий.

- Заменим все символы A в последовательности P = $\{p_k\}_{k=1}^{N}$ на 1, остальные символы — на 0. Получим первую бинарную последовательность PA = $\{p_k^{\mathsf{A}}\}_{k=1}^{N}$, соответствующую символу A в последовательности P.
- Заменим все символы C в последовательности P на 1, остальные символы — на 0. Получим вторую бинарную последовательность PC = $\{p_k^{\mathsf{C}}\}_{k=1}^{N}$, соответствующую символу C в последовательности P.
- То же самое проделаем с символами G и T в P, получим PG = $\{p_k^{\mathsf{G}}\}_{k=1}^{N}$ и PT = $\{p_k^{\mathsf{T}}\}_{k=1}^{N}$, соответственно.

- Аналогично поступим с последовательностью Q, поставив ей в соответствие последовательности QA, QC, QG и QT.

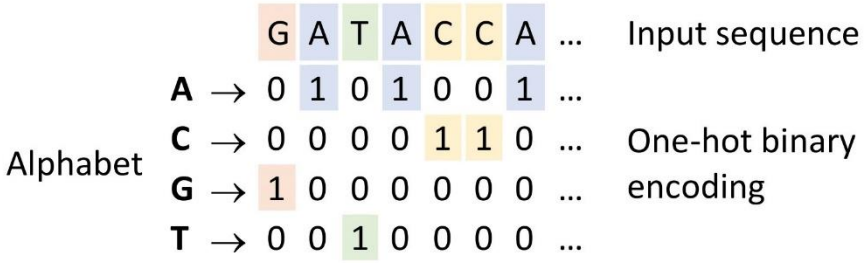

Рисунок 2. Иллюстрация преобразования символьной последовательности в набор бинарных.

Для двух бинарных последовательностей, соответствующих одному и тому же символу (например, PA $= \{p_k^{\mathrm{A}}\}_{k=1}^{N}$ и QA $= \{q_k^{\mathrm{A}}\}_{k=1}^{L}$), вычислим свёртку. Каждое значение полученной свёртки — это, по сути, количество пар {1,1} в каждом выравнивании, что в нашем случае означает, что в исходной последовательности на этих местах стоит один и тот же символ. Если для каждого символа из алфавита вычислить свёртку пары соответствующих ему бинарных последовательностей, а после просуммировать эти свёртки, то в результате получится последовательность, в которой каждое значение — это количество совпадений по всем символам. Поскольку для каждого элемента свёртки можно однозначно определить, какие «части» исходных последовательностей в это время «накладывались» друг на друга, то, следовательно мы можем определить, сколько элементов совпало в этих наложениях; следует подчеркнуть, что число точных совпадений определяется вне зависимости от того, где именно в текущем наложении встретились точные совпадения. Если число совпадений близко к общей длине пересечения, это говорит о том, что данные части последовательностей очень похожи.

Таким образом, сравнение последовательностей может быть основано на алгоритме, изложенном в следующем пункте.

2.2.2. Алгоритм метода Шайдурова

Входные данные: две последовательности P и Q длины N и L соответственно, состоящие из символов конечного алфавита א .

Одну из последовательностей необходимо инвертировать, т.е. изменить в ней порядок символов на обратный. Для определённости всегда делаем это со второй последовательностью.

Input: GGCT...CCAA, AACTC...CCGT

↓ ↓

Output: $\underbrace{\text{GGCT}\ldots\text{CCAA}}_{N}, \quad \underbrace{\text{TGCC}\ldots\text{CTCAA}}_{L}$

Каждую из последовательностей преобразуем в |א| бинарных по правилу, описанному в п. 2.2.1. Бинарные последовательности, полученные из P, обозначим Pα, α ∈ א, а полученные из Q — Qα, α ∈ א.

Поскольку применение быстрого преобразования Фурье требует использования «входной» последовательности (длина которой есть сумма длин сравниваемых последовательностей), длина которой является стпенью 2, постольку следует «входную» последовательность дополнить справа (для определённости) нулями до такой длины.

Ко всем восьми бинарным последовательностям (для случая сравнения нуклеотидных последовательностей) применяем прямое быстрое преобразование Фурье (далее — FFT). Длина последовательностей не изменяется, однако элементы становятся комплексными числами.

P˜ = {p1,p2,...,pN˜}, pi ∈ {0,1}, Q˜ = {q1,q2,...,qN˜}, qi ∈ {0,1},

⇓ ⇓

$\text{FFT}(\tilde{\mathbf{P}}) = \{p'_1, p'_2, \ldots, p'_{\tilde{N}}\},\ p'_i \in \mathbb{C}$, $\text{FFT}(\tilde{\mathbf{Q}}) = \{q'_1, q'_2, \ldots, q'_{\tilde{N}}\},\ q'_i \in \mathbb{C}$.

Для каждой пары последовательностей, соответствующих одному и тому же символу, вычисляем поэлементное произведение, затем все поэлементные произведения суммируются. Получившуюся последовательность обозначим S.

$$C: \begin{Bmatrix} p_1^C, p_2^C, \ldots, p_{\tilde{N}-1}^C, p_{\tilde{N}}^C \\ q_1^C, q_2^C, \ldots, q_{\tilde{N}-1}^C, q_{\tilde{N}}^C \end{Bmatrix} \Rightarrow \left\{ p_1^C q_1^C, p_2^C q_2^C, \ldots, p_{\tilde{N}-1}^C q_{\tilde{N}-1}^C, p_{\tilde{N}}^C q_{\tilde{N}}^C \right\}$$

$$\vdots$$

$$A: \begin{Bmatrix} p_1^A, p_2^A, \ldots, p_{\tilde{N}-1}^A, p_{\tilde{N}}^A \\ q_1^A, q_2^A, \ldots, q_{\tilde{N}-1}^A, q_{\tilde{N}}^A \end{Bmatrix} \Rightarrow \left\{ p_1^A q_1^A, p_2^A q_2^A, \ldots, p_{\tilde{N}-1}^A q_{\tilde{N}-1}^A, p_{\tilde{N}}^A q_{\tilde{N}}^A \right\},$$

,

P

S$= \left\{ p_1^A q_1^A + p_1^C q_1^C + p_1^G q_1^G + p_1^T q_1^T \right.$,

pA2 q2A + pC2 q2C + pG2 q2G + pT2q2T,...

... ,pAN˜qNA˜ + pCN˜qNC˜ + pNG˜qNG˜ + pTN˜qNT˜o.

6. К полученной сумме S применяется обратное быстрое преобразование Фурье, которое порождает числовую последовательность, в которой каждый элемент соответствует числу точных совпадений всех однородных символов.

$$S = \{s1,s2,...,sN˜\}, \quad si \downarrow \in C,$$

$$C = \{c1,c2,...,cN˜\}, \quad ci \in Z+ .$$

$$FFT{-}1(S) =$$

Следует подчеркнуть ещё раз, что это число совпадений определяется вне зависимости от того, где эти совпадения находятся в текущем наложении.

Результатом работы алгоритма, основанного на методе Шайдурова, является числовая последовательность C, каждый элемент которой соответствует количеству совпадений пар нуклеотидов в текущем наложении исходных последовательностей. Причем первый элемент **C** соответствует наложению первого символа последовательности **Q** на последний символ последовательности **P**.

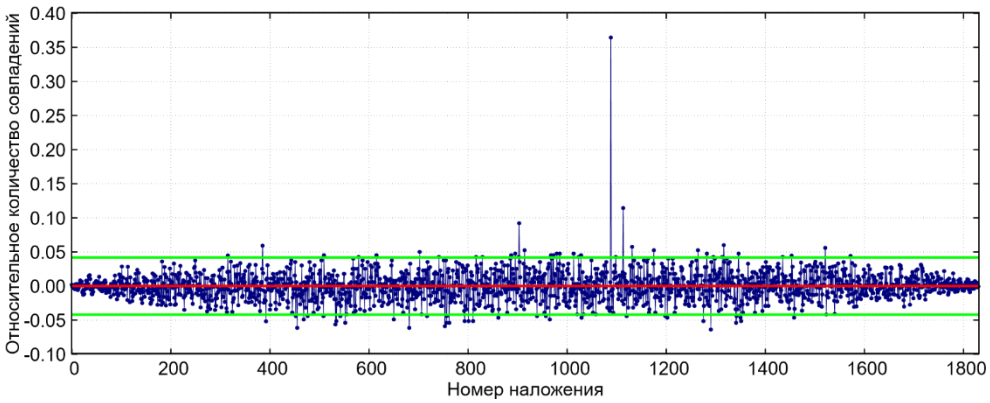

Рисунок 3. Пример карты совпадений — визуализация последовательности C.

Визуализация последовательности C представлена на рисунке 3 — это называется «картой совпадений». По оси абсцисс на ней указан порядковый номер наложения i, по оси ординат — соответствующее значение ci, нормированное на меньшую из длин сравниваемых последовательностей, именно поэтому значения свёртки не превышает единицы.

Описанный алгоритм позволяет быстро проверить пару символьных последовательностей на точное или почти точное совпадение с помощью преобразования их в бинарные.

Очевидно, что данный алгоритм не решает полностью задачу сравнения символьных последовательностей в общей постановке: во-первых, свёртка показывает только число совпадений, но не их местоположение. Действительно, одна и та же последовательность может давать одинаковое значение свёртки в разных наложениях (рис. 4).

Рисунок 4. Пример одинакового количества совпадений, равного 3-м, при различных наложениях последовательностей.

Во-вторых, одно и то же значение свертки может получиться, если две последовательности имеют общую подпоследовательность большой длины, или если две последовательности имеют много случайных совпадений малой длины (рис. 5). Такое случайное совпадение является своего рода шумом, и эта проблема требует отдельного исследования, которое не входит в задачи данной работы. На Рис. 5 показан случай совпадения четырёх символов, стоящих в разных местах наложения.

Рисунок 5. Пример одинакового количества совпадений, равного 4-м, в случае последовательностей, имеющих общую подпоследовательность большой длины, и в случае нескольких случайных совпадений (шум).

В настоящей работе было проведено сравнение выравнивания и метода Шайдурова с учётом тех трудностей, которые были выявлены к настоящему времени. Несмотря на эти трудности, использование свёрток для задач сравнения генетических последовательностей позволяет решать многие практические задачи.

Безусловным преимуществом идеи, заложенной в методе Шайдурова, по сравнению с методом выравнивания является отсутствие каких-либо свободных параметров, выбор которых произволен, что делает результаты сравнения выравниванием сильно зависящими от предпочтений конкретного исследователя. Алгоритм Шайдурова всегда детерминирован и приводит к однозначному, строго объясняемому результату.

2.3. Сравнение методом выравнивания (alignment)

Сравнение методом выравнивания было проведено с использованием инструмента попарного локального выравнивания SSEARCH2SEQ Pairwise

Sequence Alignment (PSA) [85] для последовательностей гена TRAF6. Ресурс доступен в веб-версии. Использован алгоритм локального выравнивания со стандартными параметрами штрафных функций. Для визуализации выравнивания использовалась программа BioEdit ver. 7.0.9.

2.4. Филогенетический анализ

Одной из важнейших задач выравнивания в современной биологии и генетике является установление филогенетических отношений видов или последовательностей. Филогенетический анализ основывается на процедуре множественного выравнивания (MSA), расчета расстояний между видами или последовательностями и представления этих результатов в виде древовидного графа — филогенетического дерева. При оценке результатов сверток и выравнивания были построены деревья по митохондриальным геномам 25 видов рыб.

Для выравнивания последовательностей был использован MAFFT версии 7 [86]. Для выравнивания митохондриальных геномов рыб в MAFFT наиболее подходящим алгоритмом итеративного уточнения является L-INS-i. Этот алгоритм является мощным и универсальным, особенно эффективным для выравнивания длинных последовательностей с потенциально сложной структурой, частыми вставками и делециями (инделами).

L-INS-i использует локальные выравнивания для поиска точек сопоставления и глобальные выравнивания для соединения этих точек, что обеспечивает высокую точность. Алгоритм хорошо справляется с длинными последовательностями, такими как митохондриальные геномы. Митохондриальные геномы могут содержать множество инделов, и L-INS-i эффективно справляется с их наличием.

Оптимальная модель молекулярной эволюции и филогенетическое дерево методом максимального правдоподобия были построены с использованием программы IQ-TREE версии 1.6.12. В результате была выбрана модель

TVM+I+G (Transversion Model + Invariant Sites + Gamma distribution), которая является одной из сложных моделей эволюции, используемых в филогенетических анализах для учета различных аспектов молекулярной эволюции нуклеотидных последовательностей с вариацией, включающей инвариантные сайты и гамма-распределение. Также в IQ-TREE было построено дерево методом максимального правдоподобия (Maximum Likelihood, ML). Этот метод оценивает вероятности различных гипотетических деревьев и выбирает то дерево, которое имеет наибольшую вероятность объяснить наблюдаемые данные, исходя из выбранной модели эволюции. Для визуализации деревьев использовалась программа FigTree (http://tree.bio.ed.ac.uk/software/figtree).

Построение филогенетических деревьев по результатам свёрток производилось в Python Jupyter notebook. Предлагаемый в статье [83] метод использует свертку быстрого преобразования Фурье (БПФ) для вычисления расстояния Хэмминга. Концепция построения филогенетических деревьев по результатам свёрточной функции состоит из трех этапов: предварительная обработка (кодирование) входных последовательностей, применение БПФ для свертки предварительно обработанных данных и прямая количественная интерпретация полученной свертки.

Глава 3. Результаты

ЗАКЛЮЧЕНИЕ

В данной магистерской диссертации была проведена сравнительная оценка двух методов, используемых для анализа нуклеотидных последовательностей: классического метода выравнивания и метода свёрточных функций, предложенного Шайдуровым. Основной целью работы было выявление связи между результатами резных методов при решении задач, связанных с генетическим анализом и построением филогенетических деревьев.

Классический метод выравнивания демонстрирует высокую точность при прямом сравнении последовательностей и хорошо подходит для задач, требующих детального выравнивания. Метод Шайдурова, основанный на свёрточных функциях, показал свою эффективность в задачах, связанных с обработкой больших объёмов данных и построением филогенетических деревьев, предлагая альтернативный подход к анализу. При сравнении митохондриальных геномов и последующем филогенетическом анализе были получены деревья, демонстрирующие не полностью аналогичную, но достаточно близкую топологию для метода Шайдурова и для метода выравнивания.

Сравнительный анализ гена TRAF6 показал, что результаты метода Шайдурова имеют необходимую степень корреляции с классическими методами выравнивания, что подтверждает его применимость в задачах генетического анализа, хотя с некоторыми ограничениями и особенностями. Высокая корреляция соблюдается для разных типов мутаций, однако в случае присутствия в последовательности гэпов, значения корреляции падают.

Метод свёрточных функций может быть полезен в случаях, когда требуется быстрое и менее трудоёмкое сравнение последовательностей, особенно при анализе больших генетических данных, своего рода «грубая разметка».

В заключение, данное исследование демонстрирует, что комбинированное использование обоих методов может привести к более всестороннему и

глубокому анализу нуклеотидных последовательностей, обеспечивая как точность, так и эффективность. Перспективы дальнейшего развития данного направления включают оптимизацию и интеграцию этих методов для улучшения качества и скорости генетических исследований.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Basic local alignment search tool / Stephen F Altschul, Warren Gish, Webb Miller et al. // Journal of molecular biology. — 1990. — Vol. 215, no. 3. — Pp. 403–410.
2. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positionspecific gap penalties and weight matrix choice / Julie D Thompson, Desmond G Higgins, Toby J Gibson // Nucleic acids research. — 1994. — Vol. 22, no. 22. — Pp. 4673–4680.
3. Alignment-free inference of hierarchical and reticulate phylogenomic relationships / Guillaume Bernard, Cheong Xin Chan, Yao-ban Chan et al. // Briefings in Bioinformatics. — 2019. — Vol. 20, no. 2. — Pp. 426–435.
4. Alignment-free sequence comparison: benefits, applications, and tools / Andrzej Zielezinski, Susana Vinga, Jonas Almeida, Wojciech M Karlowski // Genome biology. — 2017. — Vol. 18. — Pp. 1–17.
5. A statistical method for alignment-free comparison of regulatory sequences / Miriam R Kantorovitz, Gene E Robinson, Saurabh Sinha // Bioinformatics. — 2007. — Vol. 23, no. 13. — Pp. i249– i255.
6. Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs / Andra Ivan, Marc S Halfon, Saurabh Sinha // Genome Biology. — 2008. — Vol. 9. — Pp. 1–17.
7. Rapid similarity search of proteins using alignments of domain arrangements / Nicolas Terrapon, January Weiner, Sonja Grath et al. // Bioinformatics. — 2014. — Vol. 30, no. 2. — Pp. 274–281.
8. A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF / Yingnan Cong, Yao-ban Chan, Mark A Ragan // Scientific reports. — 2016. — Vol. 6, no. 1. — P. 30308.
9. Mash: fast genome and metagenome distance estimation using MinHash / Brian D Ondov, Todd J Treangen, Pa´ll Melsted et al. // Genome biology — 2016. — Vol. 17. — Pp. 1–14.

10. Classification of methanogenic bacteria by 16S ribosomal RNA characterization / George E Fox, Linda J Magrum, William E Balch et al. // Proceedings of the National Academy of Sciences. — 1977. — Vol. 74, no. 10. — Pp. 4537–4541.
11. Alignment-free sequence comparison—a review / Susana Vinga, Jonas Almeida // Bioinformatics. — 2003. — Vol. 19, no. 4. — Pp. 513–523.
12. A general method applicable to the search for similarities in the amino acid sequence of two proteins / Saul B Needleman, Christian D Wunsch // Journal of molecular biology. — 1970. — Vol. 48, no. 3. — Pp. 443–453.
13. Waterman, Michael S. Some biological sequence metrics / Michael S Waterman, Temple F Smith, William A Beyer // Advances in Mathematics. — 1976. — Vol. 20, no. 3. — Pp. 367–387.
14. Duffy, Siobain. Rates of evolutionary change in viruses: patterns and determinants / Siobain Duffy, Laura A Shackelton, Edward C Holmes // Nature Reviews Genetics. — 2008. — Vol. 9, no. 4. — Pp. 267–276.
15. Xiong, Jin. Essential bioinformatics / Jin Xiong. // Cambridge University Press — 2006.
16. Twilight zone of protein sequence alignments / Burkhard Rost // Protein engineering. — 1999. — Vol. 12, no. 2. — Pp. 85–94.
17. A statistical physics perspective on alignmentindependent protein sequence comparison / Amit K Chattopadhyay, Diar Nasiev, Darren R Flower // Bioinformatics. — 2015. — Vol. 31, no. 15. — Pp. 2469–2474.
18. Quantifying the relationship between sequence and three-dimensional structure conservation in RNA / Emidio Capriotti, Marc A Marti-Renom // BMC bioinformatics. — 2010. — Vol. 11. — Pp. 1–10.
19. What is dynamic programming? / Sean R Eddy // Nature biotechnology. — 2004. — Vol. 22, no. 7. — Pp. 909–910.
20. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement / Aaron E Darling, Bob Mau, Nicole T Perna // PloS one. — 2010. — Vol. 5, no. 6. — P. e11147.

21. Mugsy: fast multiple alignment of closely related whole genomes / Samuel V Angiuoli, Steven L Salzberg // Bioinformatics. — 2011. — Vol. 27, no. 3. — Pp. 334–342.
22. Multiple sequence alignment modeling: methods and applications / Maria Chatzou, Cedrik Magis, Jia-Ming Chang et al. // Briefings in bioinformatics. — 2016. — Vol. 17, no. 6. — Pp. 1009–1023.
23. Alignment uncertainty and genomic analysis / Karen M Wong, Marc A Suchard, John P Huelsenbeck // Science. — 2008. — Vol. 319, no. 5862. — Pp. 473–476.
24. Clustal W and Clustal X version 2.0 / Mark A Larkin, Gordon Blackshields, Nigel P Brown et al. // bioinformatics. — 2007. — Vol. 23, no. 21. — Pp. 2947–2948.
25. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs / Stephen F Altschul, Thomas L Madden, Alejandro A Scha¨ffer et al. // Nucleic acids research. — 1997. — Vol. 25, no. 17. — Pp. 3389–3402.
26. Grid-enabled blastz: application to comparative genomics / Chunxi Chen, Jagath C Rajapakse // The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology. — 2007. — Vol. 48. — Pp. 301–309.
27. BLAT—the BLAST-like alignment tool / W James Kent // Genome research. — 2002. — Vol. 12, no. 4. — Pp. 656–664.
28. Computational solutions to large-scale data management and analysis / Eric E Schadt, Michael D Linderman, Jon Sorenson et al. // Nature reviews genetics. — 2010. — Vol. 11, no. 9. — Pp. 647–657.
29. Multiple sequence alignment with the Clustal series of programs / Ramu Chenna, Hideaki Sugawara, Tadashi Koike et al. // Nucleic acids research. — 2003. — Vol. 31, no. 13. — Pp. 3497–3500.
30. Parallelization of the MAFFT multiple sequence alignment program / Kazutaka Katoh, Hiroyuki Toh // Bioinformatics. — 2010. — Vol. 26, no. 15. — Pp. 1899–1900.

31. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution / Se-Ran Jun, Gregory E Sims, Guohong A Wu, Sung-Hou Kim // Proceedings of the National Academy of Sciences. — 2010. — Vol. 107, no. 1. — Pp. 133–138.
32. Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison / Chris-Andre Leimeister, Burkhard Morgenstern // Bioinformatics. — 2014. — Vol. 30, no. 14. — Pp. 2000–2008.
33. An estimator for local analysis of genome based on the minimal absent word / Lianping Yang, Xiangde Zhang, Haoyue Fu, Chenhui Yang // Journal of Theoretical Biology. — 2016. — Vol. 395. — Pp. 23–30.
34. Alignment free comparison: similarity distribution between the DNA primary sequences based on the shortest absent word / Lianping Yang, Xiangde Zhang, Hegui Zhu // Journal of theoretical biology. — 2012. — Vol. 295. — Pp. 125–131.
35. Fast and accurate estimation of evolutionary distances between closely related genomes / Bernhard Haubold, Fabian Klotzl, Peter Pfaffelhuber // Bioinformatics. — 2015. — Vol. 31, no. 8. — Pp. 1169– 1175.
36. ‘Multi-SpaM’: a maximum-likelihood approach to phylogeny reconstruction using multiple spaced-word matches and quartet trees / Thomas Dencker, Chris-Andr´e Leimeister, Michael Gerth et al. // NAR Genomics and Bioinformatics. — 2020. — Vol. 2, no. 1. — P. lqz013.
37. Fast and accurate phylogeny reconstruction using filtered spaced-word matches / Chris-Andre Leimeister, Salma SohrabiJahromi, Burkhard Morgenstern // Bioinformatics. — 2017. — Vol. 33, no. 7. — Pp. 971–979.
38. Prot-SpaM: Fast alignment-free phylogeny reconstruction based on wholeproteome sequences / Chris-Andre Leimeister, Jendrik Schellhorn, Svenja Do¨rrer et al. // GigaScience. — 2019. — Vol. 8, no. 3. — P. giy148.
39. A protein map and its application / Stephen S-T Yau, Chenglong Yu, Rong He // DNA and cell biology. — 2008. — Vol. 27, no. 5. — Pp. 241–250.

40. An improved model for whole genome phylogenetic analysis by Fourier transform / Changchuan Yin, Stephen S-T Yau // Journal of theoretical biology. — 2015. — Vol. 382. — Pp. 99–110.
41. Information theory applications for biological sequence analysis / Susana Vinga // Briefings in bioinformatics. — 2014. — Vol. 15, no. 3. — Pp. 376–389.
42. Sequence analysis by iterated maps, a review / Jonas S Almeida // Briefings in bioinformatics. — 2014. — Vol. 15, no. 3. — Pp. 369–375.
43. A survey and evaluations of histogram-based statistics in alignment-free sequence comparison / Brian B Luczak, Benjamin T James, Hani Z Girgis // Briefings in bioinformatics. — 2019. — Vol. 20, no. 4. — Pp. 1222–1237.
44. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions / Gregory E Sims, Se-Ran Jun, Guohong A Wu, Sung-Hou Kim // Proceedings of the National Academy of Sciences. — 2009. — Vol. 106, no. 8. — Pp. 2677–2682.
45. CAFE: a C celerated A lignment-F r E e sequence analysis / Yang Young Lu, Kujin Tang, Jie Ren et al. // Nucleic acids research. — 2017. — Vol. 45, no. W1. — Pp. W554–W559.
46. Inferring phylogenies of evolving sequences without multiple sequence alignment / Cheong Xin Chan, Guillaume Bernard, Olivier Poirion et al. // Scientific reports. — 2014. — Vol. 4, no. 1. — P. 6504.
47. Alignment-free sequence analysis and applications / Jie Ren, Xin Bai, Yang Young Lu et al. // Annual Review of Biomedical Data Science. — 2018. — Vol. 1. — Pp. 93–114.
48. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis / Oliver Bonham-Carter, Joe Steele, Dhundy Bastola // Briefings in bioinformatics. — 2014. — Vol. 15, no. 6. — Pp. 890–905.
49. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing / Kai Song, Jie Ren, Gesine Reinert

et al. // Briefings in bioinformatics. — 2014. — Vol. 15, no. 3. — Pp. 343–353.

50. Abundance of ultramicro inversions within local alignments between human and chimpanzee genomes / Yuichiro Hara, Tadashi Imanishi // BMC evolutionary biology. — 2011. — Vol. 11. — Pp. 1–11.

51. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer / Guillaume Bernard, Cheong Xin Chan, Mark A Ragan // Scientific reports. — 2016. — Vol. 6, no. 1. — P. 28970.

52. Alignment-free methods in computational biology. — 2014. The average common substring approach to phylogenomic reconstruction / Igor Ulitsky, David Burstein, Tamir Tuller, Benny Chor // Journal of Computational Biology. — 2006. — Vol. 13, no. 2. — Pp. 336–350.

53. Genome comparison without alignment using shortest unique substrings / Bernhard Haubold, Nora Pierstorff, Friedrich Moller, Thomas Wiehe // BMC bioinformatics. — 2005. — Vol. 6. — Pp. 1–11.

54. On finding minimal absent words / Armando J Pinho, Paulo JSG Ferreira, Sara P Garcia, Jo˜ao MOS Rodrigues // BMC bioinformatics. — 2009. — Vol. 10. — Pp. 1–11.

55. Large local analysis of the unaligned genome and its application / Lianping Yang, Xiangde Zhang, Tianming Wang, Hegui Zhu // Journal of Computational Biology. — 2013. — Vol. 20, no. 1. — Pp. 19–29.

56. A 2D graphical representation of protein sequence and its numerical characterization / Jia Wen, YuYan Zhang // Chemical Physics Letters — 2009. — Vol. 476, no. 4-6. — Pp. 281–286.

57. MEGA-core of phylogenetic analysis in molecular evolutionary genetics / N Khan // J Phylogen Evol Biol. — 2017. — Vol. 5, no. 2. — P. 1000183.

58. Benchmarking of alignment-free sequence comparison methods / Andrzej Zielezinski, Hani Z Girgis, Guillaume Bernard et al. // Genome biology. — 2019. — Vol. 20. — Pp. 1–18.

59. The organization and inheritance of the mitochondrial genome / Xin Jie Chen, Ronald A Butow // Nature Reviews Genetics. — 2005. — Vol. 6, no. 11. — Pp. 815–825.
60. Modifying the mitochondrial genome / Alexander N Patananan, TingHsiang Wu, Pei-Yu Chiou, Michael A Teitell // Cell metabolism. — 2016. — Vol. 23, no. 5. — Pp. 785–796.
61. Mitochondrial genomic landscape: a portrait of the mitochondrial genome 40 years after the first complete sequence / Alessandro Formaggioni, Andrea Luchetti, Federico Plazzi // Life. — 2021. — Vol. 11, no. 7. — P. 663.
62. A comprehensive analysis of bilaterian mitochondrial genomes and phylogeny / Matthias Bernt, Christoph Bleidorn, Anke Braband et al. // Molecular Phylogenetics and Evolution. — 2013. — Vol. 69, no. 2. — Pp. 352–364.
63. Mitochondrial genome function and maternal inheritance / Allen J. F., de Paula W. B. M. // 2013 — 1298-1304.
64. Structure and variation of the mitochondrial genome of fishes / Takashi P Satoh, Masaki Miya, Kohji Mabuchi, Mutsumi Nishida // BMC genomics. — 2016. — Vol. 17. — Pp. 1–20.
65. Extent and scale of local adaptation in salmonid fishes: review and metaanalysis / Dylan J Fraser, Laura K Weir, Louis Bernatchez et al. // Heredity. — 2011. — Vol. 106, no. 3. — Pp. 404–420.
66. The complete sequence of the zebrafish (Danio rerio) mitochondrial genome and evolutionary patterns in vertebrate mitochondrial DNA / Richard E Broughton, Jami E Milam, Bruce A Roe. // Genome research. — 2001. — Vol. 11, no. 11. — Pp. 1958–1967.
67. Mitochondrial genome (mtDNA) mutations that generate reactive oxygen species / Anne Hahn, Steven Zuryn // Antioxidants. — 2019. — Vol. 8, no. 9. — P. 392.
68. Patterns of natural selection acting on the mitochondrial genome of a locally adapted fish species / Sofia Consuegra, Elgan John, Eric Verspoor, Carlos

Garcia de Leaniz // Genetics Selection Evolution. — 2015. — Vol. 47. — Pp. 1–10.

69. Fish mitochondrial genome sequencing: expanding genetic resources to support species detection and biodiversity monitoring using environmental DNA / Julie C Schroeter, Aaron P Maloy, Christopher B Rees, Meredith L Bartron // Conservation Genetics Resources. — 2020. — Vol. 12, no. 3. — Pp. 433–446.

70. MitoFish and MiFish pipeline: a mitochondrial genome database of fish with an analysis pipeline for environmental DNA metabarcoding / Yukuto Sato,

71. Molecular biology and evolution / Masaki Miya, Tsukasa Fukunaga et al. // — 2018. — Vol. 35, no. 6. — Pp. 1553–1555.

72. Mitochondrial genome structure and composition in 70 fishes: a key resource for fisheries management in the South Atlantic / Marcela Alvarenga, Ananda Krishna Pereira D'Elia, Graciane Rocha et al. // BMC genomics. — 2024. — Vol. 25, no. 1. — P. 215.

73. TNF receptor associated factors (TRAFs) / Hao Wu. //— Springer Science & Business Media, 2007. — Vol. 597.

74. The ubiquitin E3 ligase TRAF6 exacerbates pathological cardiac hypertrophy via TAK1-dependent signalling / Yan-Xiao Ji, Peng Zhang, XiaoJing Zhang et al. // Nature communications. — 2016. — Vol. 7, no. 1. — P. 11267.

75. Protein-binding function of RNA-dependent protein kinase promotes proliferation through TRAF2/RIP1/NF-κB/c-Myc pathway in pancreatic β cells / LiLi Gao, Wei Tang, ZhengZheng Ding et al. // Molecular Medicine. — 2015. — Vol. 21. — Pp. 154–166.

76. Enhancement of Toll-like receptor3 (TLR3)-induced death signaling by TNF-like weak inducer of apoptosis (TWEAK): Ph.D. thesis / Anany, Mohamed Ahmed Mohamed Mohamed. // Universitat Wurzburg. — 2019.

77. TGFβ-induced Degradation of TNF Receptor-associated Factor 3 (TRAF3) in Mesenchymal Progenitor Cells Causes Age-related Osteoporosis: Ph.D. thesis / Li, Jinbo. // University of Rochester. — 2020.

78. TRAF6 neddylation drives inflammatory arthritis by increasing NF-κB activation / Kewei Liu, Kaizhe Chen, Qian Zhang et al. // Laboratory Investigation. — 2019. — Vol. 99, no. 4. — Pp. 528–538.
79. Increased alveolar epithelial TRAF6 via autophagy-dependent TRIM37 degradation mediates particulate matter-induced lung metastasis / Jiajun Liu, Shumin Li, Xuefeng Fei et al. // Autophagy. — 2022. — Vol. 18, no. 5. — Pp. 971–989.
80. Recognition of TRAIP with TRAFs: current understanding and associated diseases / Nasreena Sajjad, Mohammad Muzaffar Mir, Johra Khan et al. // The International Journal of Biochemistry & Cell Biology. — 2019. — Vol. 115. — P. 105589.
81. Complete mitochondrial DNA sequences of the threadfin cichlid (Petrochromis trewavasae) and the blunthead cichlid (Tropheus moorii) and patterns of mitochondrial genome evolution in cichlid fishes / Christoph Fischer, Stephan Koblmuller, Christian Gully et al. // PLoS One. — 2013. — Vol. 8, no. 6. — P. e67048.
82. Highly parallel convolution method to compare DNA sequences with enforced in/del and mutation tolerance / Anna Molyavko, Vladimir Shaidurov, Eugenia Karepova, Michael Sadovsky // Bioinformatics and Biomedical Engineering: 8th International Work-Conference, IWBBIO 2020, Granada, Spain, May 6–8, 2020, Proceedings 8 / Springer. — 2020. — Pp. 472–481.
83. Efficient Clustering of Genetic Entities / Anna Molyavko, Igor Borovikov, Evgenia Karepova, Michael Sadovsky // 2022 12th International Conference on Pattern Recognition Systems (ICPRS) / IEEE. — 2022. — Pp. 1–6.
84. Comparison of the Alignment and Shaidurov's Methods the Search Efficiency in Symbol Sequences with Mismatches / Anna Molyavko, Evgenia Karepova, Mikhail Sadovsky et al. // CEUR Workshop Proceedings. — 2021. — Pp. 93–97.

85. Search and sequence analysis tools services from EMBL-EBI in 2022 / Fabio Madeira, Matt Pearce, Adrian RN Tivey et al. // Nucleic acids research. — 2022. — Vol. 50, no. W1. — Pp. W276–W279.
86. Katoh, Kazutaka. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization / Kazutaka Katoh, John Rozewicki, Kazunori D Yamada // Briefings in bioinformatics. — 2019. — Vol. 20, no. 4. — Pp. 1160–1166.

Министерство науки и высшего образования РФ
Федеральное государственное автономное
образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт фундаментальной биологии и биотехнологии

Кафедра геномики и биоинформатики

УТВЕРЖДАЮ

Заведующий кафедрой

_____И.Е. Ямских

«20» июня 2024 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Два метода сравнения нуклеотидных последовательностей – выравнивания (alignment) и метода Шайдурова, проблемы и перспективы

06.04.01 – Биология

06.04.01.06 – Геномика и биоинформатика

Руководитель	_____ подпись, дата	д.ф.-м.н., проф. должность, ученая степень	М.Г. Садовский инициалы, фамилия
Выпускник	_____ подпись, дата		А.А. Тетерлева инициалы, фамилия
Рецензент	_____ подпись, дата	д.ф.-м.н. ученая степень	С.И. Барцев инициалы, фамилия

Красноярск 2024