

Санкт-Петербургский Государственный Университет
Математико-механический факультет

Кафедра Системного Программирования

Крамар Алина Сергеевна

Построение генетических карт по неполным и зашумленным данным

Дипломная работа

Допущена к защите.
Зав. кафедрой:
д. ф.-м. н., профессор Терехов А. Н.

Научный руководитель:
к. ф.-м. н., доцент Сысоев С. С.

Рецензент:
науч. сотрудник. Добрынин П.В.

Санкт-Петербург
2014

Saint Petersburg State University
Mathematics & Mechanics Faculty

Department of Software Engineering

Alina Kramar

Genetic linkage mapping based on incomplete and noisy data

Graduation Thesis

Admitted for defence.

Head of the chair:
professor Andrey Terekhov

Scientific supervisor:
associate professor Sergey Sysoev

Reviewer:
research associate Pavel Dobrynin

Saint-Petersburg
2014

Оглавление

Введение	4
1. Обзор существующих решений	7
1.1. Алгоритм Элстона-Стюарта	8
1.2. Алгоритм Ландера-Грина	9
1.3. Построение генетических карт по полностью секвенированным участкам геномов	10
2. Постановка задачи	12
3. Доработка алгоритма	13
3.1. Анализ недостатков и их исправления	13
3.1.1. Восстановление расположения маркеров по матрице попарных расстояний	13
3.1.2. Повышение точности матрицы частот рекомбинаций	17
3.2. Доработанный алгоритм	17
4. Сравнение алгоритмов	19
4.1. Генерация тестовых данных	20
4.2. Процесс сравнения и анализа	20
4.3. Обобщённые данные	22
Заключение	24
Список литературы	25

Введение

Краткая медицинская энциклопедия [1] даёт следующее определение слову “генетика”: наука о наследственности и изменчивости организма. Согласно законам наследования все основные признаки и свойства любых организмов определяются и контролируются единицами наследственной информации - генами, локализованными в специфических структурах клетки - хромосомах [2, 3]. В связи с этим основной задачей генетики является на основе первичной структуры биополимеров (молекул ДНК или РНК) определить фенотип особи, механизмы наследования тех или иных признаков и т.п. Получение информации о первичной структуре ДНК называется секвенированием [4]. К сожалению, современные методы секвенирования не в состоянии предоставить информацию о полной нуклеотидной последовательности в рамках конкретной хромосомы [5]. Сама по себе нуклеотидная последовательность не содержит прямой информации о происхождении того или иного однозначно идентифицируемого участка (маркера), и поэтому довольно сложно определить, от какого предка был унаследован тот или иной признак. Для определения шаблонов наследования и выявления его принципов используются генетические карты хромосом [6].

Генетическая карта - это схема или порядок расположения маркеров на хромосоме (здесь и далее подразумевается структурный маркер, т.е. маркер, имеющий отличное нуклеотидное представление, которое позволяет его идентифицировать) и генов. Зачастую, наличие генов не так важно, как наличие маркеров, потому что гены имеют свойство наследоваться не полностью [7]. Идея создать генные карты принадлежит Томасу Моргану, внёсшему огромный вклад в теорию наследственности. Его идея была связана с явлением сцепленного наследования генов. Из-за мейотического кроссинговера, который делает невозможным полностью коррелированное наследование маркеров и влияет на расхождение сцепленных генов по разным гаметам, у Морганна появилось предположение о существовании связи между физическим расстоянием между маркерами и их взаимодействием в процессе наследования [8]. Это предположение подтвердилось в 1913 году учеником Морганна Альфредом Стёртевантом, построившим первую генетическую карту на основе данных *Drosophila melanogaster* [9, 10].

Кроссинговер на этапе мейоза вносит возмущение в сцепленное наследование [8], и чем чаще он проявляется, тем чаще наблюдается отклонение от сцеплений. Физическое расстояние между парой генов прямо пропорционально вероятности кроссинговера, что позволяет однозначно расположить маркеры на молекуле [10]. Генетическое расстояние не трудно перевести в физическое, но чаще всего в этом нет необходимости, так как генетическая карта несёт структурный характер и для механизма наследования существенен порядок [11, 12] исследуемых маркеров. Единицей измерения расстояния является 1 сантиморган (1 сМ). Стоит упомянуть, что задача построе-

ния генетической карты хромосомы имеет смысл для диплоидных особей и лучше решается, когда исследуемые особи находятся в родстве [13]. Задача так же решается проще с помощью данных о семействах особей, которые размножаются с большой скоростью (Быстрое размножение и большое количество потомков позволяют секвенировать большее количество особей из разных поколений за меньший период, поэтому данные о таких родословных получаются наиболее информативными), поэтому большее количество карт на данный момент имеют хромосомы дрозофил, кошек, мелких грызунов и насекомых [14]. В данном исследовании нас интересуют генетические карты человека, так как эти карты являются единственным способом проведения генетического анализа на наличие и предрасположенность особи к тяжёлым наследственным заболеваниям. С помощью генетического анализа можно выявлять болезнь Альцгеймера [15], гемофилию [16], хорею Хаттингтона [17] и т.п.

Современные средства генетического картирования позволяют построить карту по файлу с данными о родословной в определённом формате [18, 19, 20]. Причём не обязательно иметь полную информацию о степени родства, половой принадлежности и возрасте рассматриваемых особей. Более того, результаты секвенирования могут содержать ошибки.

Общепризнанными [21] и наиболее распространёнными методами генетического картирования являются:

- Алгоритм Элстона-Стюарта
- Алгоритм Ландера-Грина

Программные средства, основанные на вышеуказанных алгоритмах, хорошо решают поставленную задачу на взятом у человеческих особей материале при сравнительно небольшом (по сравнению с количеством известных науке видами маркеров) количестве маркеров [Алгоритм Элстона—Стюарта] и исследуемой семьи для небольшого размера [Алгоритм Ландера—Грина]. Практические требования современной медицины приводят к необходимости строить генетические карты по все большему и большему наборам маркеров. Вычислительная сложность алгоритма Элстона-Стюарта зависит от количества исследуемых маркеров экспоненциально [22], поэтому время работы алгоритма растёт с увеличением данного параметра. В результате чего генетическое картирование этим методом становится неосуществимым на практике. Алгоритм Ландера-Грина позволяет исследовать большее количество маркеров, но на данных о семействах меньшего размера, так как время работы этого алгоритма экспоненциально растёт с увеличением количества наблюдаемых в родословной особей—“непрародителей” (особей, родители которых присутствуют в исследуемом множестве особей) [22].

В 2013 году в работе [23] Сергеем Сысоевым был предложен алгоритм построения генетической карты путём прямого извлечения информации о генетических рассто-

яниях между маркерами без учёта кратности кроссинговера. Вычислительная сложность данного алгоритма имеет полиномиальную зависимость от количества маркеров, так и от мощности множества особей в рассматриваемой родословной, что делает данный алгоритм перспективнее ранее упомянутых для решения задач современной генетики. При этом алгоритм прямого извлечения данных имеет ряд существенных недостатков, а именно:

1. Плохая теоретическая обоснованность.

В работе [23] Сысоевым утверждается, что алгоритм извлекает всю возможную информацию о рекомбинациях маркеров в исследуемых особях. В главе 1.3 мы покажем, что это утверждение вообще говоря неверно.

2. Отсутствие верификации.

Алгоритм был проверен только на небольшой семье кошек из 192 особей и рассматривал 35 маркеров. На других данных алгоритм не проверялся.

3. Неверный результат в случае кратности кроссинговера.

Это существенно снижает полезность алгоритма, так как в природе неоднократный кроссинговер — явление, встречающееся достаточно часто.

Исходя из этого, цель данной работы можно сформулировать следующим образом: предложить новый алгоритм построения генетических карт на основе предложенного в статье [23] метода, усовершенствовав алгоритм в части учёта кратных рекомбинаций, верифицируя полученный алгоритм на реальных и синтетических данных, сравнив результаты работы нового алгоритма с результатами и характеристиками работы аналогов.

1. Обзор существующих решений

Основные понятия

Для того, чтобы рассматривать существующие подходы, нужно более формально поставить задачу, которую они решают. Нам потребуются следующие понятия [24]:

Маркер(ДНК-маркер) — полиморфный признак, выявляемый на уровне нуклеотидной последовательности ДНК.

Локус — положение маркера на генетической или цитологической карте.

Аллель — вариант последовательности ДНК в текущем локусе.

Гомозигота — диплоидная (двойной набор одинаковых хромосом) особь, копия генов которой представлена одинаковыми аллелями.

Гетерозигота — диплоидная особь, копия генов которой представлена разными аллелями.

Поясним введенные выше понятия примером. Информацию об особи будем записывать в виде строки вида **AaBb**, где **A** — вид маркера в отцовской хромосоме в позиции 1, **a** — вид маркера в материнской хромосоме в позиции 1, **B** — вид маркера в отцовской хромосоме в позиции 2, **b** — вид маркера в материнской хромосоме в позиции 2. В данном примере особь гетерозиготна в позициях 1 и 2, так как имеет разные маркеры от отца и матери. Если бы в позиции 2 находилась строка **BB**, то особь была бы гомозиготна, так как маркеры одинаковые.

Гамета — одинарный набор хромосом. В нашем случае, подстрока последовательности аллелей, содержащая половину генетического материала.

Фаза — различают две фазы: CIS и TRANS. CIS — расположение доминантных генов на одной хромосоме, а TRANS — расположение доминантных генов на разных хромосомах. Фаза имеет смысл только для особей, гетерозиготных в обоих локусах.

В нашем примере особь **AaBb** находится в фазе CIS, так как доминантные признаки (символы в верхнем регистре) находятся на одной хромосоме. **AabB** — пример TRANS фазы.

Рекомбинация — перераспределение генетического материала родителей в потомстве.

В приведённом выше примере при образовании особью гамет, которая передаётся потомкам, могут возникнуть следующие сочетания: **AB**, **Ab**, **aB** и **ab**. Получить информацию о наличии рекомбинации мы можем, зная фазу. Для CIS-фазы (в нашем случае, особь представляется строкой **AaBb** или **aAbB**), рекомбинантными будут являться гаметы **aB** и **Ab**, в случае же TRANS-фазы (особь — строка вида **AabB** или **aABb**), рекомбинантными являются гаметы **ab** и **AB**.

Генетическая карта — схема расположения маркеров на хромосоме.

Кроссинговер(Перекрёст) — процесс обмена участками гомологичных хромосом.

Перейдём к алгоритмическому смыслу задачи и рассмотрим пути её решения. Имея на входе данные о родословной, количестве исследуемых маркеров, о нуклеотидных последовательностях всех особей, а так же их родственной связи, генетическую карту можно построить с помощью следующих алгоритмов, имеющими программную реализацию [22]:

- Алгоритм Элстона - Стюарта
- Алгоритм Ландера - Грина
- Построение генетических карт по полностью секвенированным участкам геномов (Сысоев)

Алгоритм Элстона-Стюарта, так же как и алгоритм Ландера-Грина, трудно назвать алгоритмами в классическом понимании этого слова, так как они не предоставляют собой рабочего механизма, приводящего сразу к ответу, а скорее позволяют привести задачу в биологической и формальной формулировке в задачу в математической записи, которая хорошо изучена, имеет решения и хорошо моделируется. Так, например, алгоритм Ландера-Грина формулирует задачу генетического картирования в терминах скрытых марковских цепей, а алгоритм Элстона-Стюарта оптимизирует функционал правдоподобия для конкретной родословной, упрощая его вычисления.

1.1. Алгоритм Элстона-Стюарта

Нам известно, что генотип родителя влияет на генотип ребёнка, поэтому можем записать вероятность генотипа потомка как

$$P(G_{founder}).$$

Так же мы знаем, что вероятность генотипа потомка контролируется родительскими генотипами

$$P(G_{offspring}|G_{mother}, G_{father}).$$

Алгоритм Элстона-Стюарта, как и алгоритм Ландера-Грина, не связывает частоту рекомбинаций с расстоянием между генами и не рассматривает влияние расстояния на сцепление при наследовании, что в нашем случае означает, что вышеуказанная изначально никак не параметризуется.

Вероятность наблюдаемого фенотипа контролируется генотипом

$$P(X_i|G_i).$$

Рассмотрим функционал правдоподобия для задачи генетического картирования в общем случае, используя записанные выше вероятности [25]:

$$L = \Sigma_{G_1} \dots \Sigma_{G_n} P_{i_{\{o,f,m\}}} P(G_o|G_f, G_m) P_{i_i} P(X_i|G_i)$$

Если обратить внимание на запись вероятностей родительского генотипа, потомка данного родителя и вероятность фенотипа данного ребёнка, в попробовать посчитать их в случае родословных без кровосмесительных “браков” и содержащих не более чем два поколения (иначе говоря, каждая особь из родословной является либо родителем, либо ребёнком), то можно заметить, что большинство вероятностей оборачивается в ноль (так как нет зависимости).

Используя это наблюдение, можно записать новый функционал правдоподобия, актуальных для семейств из двух поколений.

$$L = \Sigma_{G_{founder}} P(X_{founder}|G_{founder}) H(G_{founder}) P(G_{founder})$$

— где $H_j(G_i)$ — вероятность наличия у G_j детей вместе с G_i . При максимизации данного функционала, просмотр особей начинается с нижнего уровня, из-за обнуления большей части вероятностей.

Нетрудно увидеть, что максимизация описанного функционала правдоподобия влечёт большую вычислительную сложность, а именно экспоненциальный рост времени выполнения с увеличением количества исследуемых маркеров [22]. Плюсами данного алгоритма является хорошая применимость в случае потенциально больших семейств, или в случае конечного и малого количества исследуемых маркеров.

Данный алгоритм имеет не одну программную реализацию, но для сравнения алгоритмов в конце данной работы мы будем использовать пакет FASTLINK [20], так как это самая последняя версия алгоритма, а значит и наиболее надёжная.

1.2. Алгоритм Ландера-Грина

В отличие от алгоритма Элстона-Стюарта, алгоритм Ландера-Грина основан на другом предположении о вероятных рекомбинациях. а именно, строит скрытую мар-

ковскую цепь [26], основываясь на вероятностях рекомбинации в позиции n , от рекомбинации в позиции $n - 1$. Если не вдаваться в подробности, то мы, чтобы получить данные о той или иной позиции должны рассмотреть все предшествующие маркеры, а так же рассмотреть последующие.

Запишем рассматриваемые генотипы в виде X_i , а состояние идентичных по происхождению аллелей в каждом маркере I_i $i \in (1 : M)$. Запишем вероятность генотипа, проявляемого из-за того или иного состояния маркера

$$P(X_i|I_i)$$

и воспользуемся предположением, что вероятность рекомбинации в позиции n обусловлена вероятностью рекомбинации в позиции $n - 1$, запишем вероятность

$$P(I_{i-1}|I_i)$$

получаемый функционал качества записывается следующим образом:

$$L = \sum_{I_1} \dots \sum_{I_M} P(I_1) \prod_{i=2}^M P(I_i|I_{i-1}) \prod_{i=1}^M P(X_i|I_i)$$

Нетрудно заметить, что недостатком этого подхода является экспоненциальная вычислительная сложность в зависимости от объёма родословной [27, 22]. Т.о. при больших семействах этот алгоритм неэффективен и рискует не завершиться. Зато он позволяет исследовать любое число маркеров, что делает его более удобным для задач современной генетики, так как геномы становятся сложнее, а родословные человека не настолько большие. Зато данный алгоритм позволяет исследовать семьи с кровосмесительным спариванием, как, например, происходит у кошек, мелких грызунов и т.п.

У данного алгоритма так же есть реализованное программное средство CRIMAP [19], которое будет принимать участие в сравнении алгоритмов генетического картирования.

1.3. Построение генетических карт по полностью секвенированным участкам геномов

Предыдущие алгоритмы рассматривали процесс передачи генов как лишь вероятностный процесс, не строя предположений о том, как происходит наследование. Использовать в качестве данных о происхождении аллели данные из прямых потомков и предков на основе результатов их секвенирования. Алгоритм был предложен в 2013 году и базировался на знании о том, что частота рекомбинаций между двумя маркерами прямо пропорционально расстоянию между ними [23]. В генетическом картировании появилась новая идея не максимизировать функционал качества, а смотреть

на структурное представление ДНК и извлекать информацию из него. Так же в статье [23] продемонстрирована лучшая вычислительная сложность, которая является полиномиальной, а так же хорошая точность в случае с реальными данными кошек, которые, как известно, не укладываются в область применения алгоритма Элстона-Стюарта.

Недостатком данного алгоритма является его неопробованность, зато к уверенным достоинствам можно отнести полиномиальную сложность и удобство использования. Данный алгоритм имеет программное средство в виде скрипта на языке Python2 и будет использовано нами в качестве базы для собственной реализации.

Построение генетических карт

Поскольку существующие алгоритмы имеют большую вычислительную сложность или не способны давать правильный ответ в наиболее частых случаях, было решено написать новый алгоритм, позволяющий строить генетические карты быстрее и качественнее, чем у существующих. Так как в случае с экспоненциальным ростом (алгоритм Элстона-Стюарта и алгоритм Ландера-Грина) помочь могут только локальные оптимизации [28, 22, 29], не влияющие на асимптотическую скорость, разумно предположить, что развивать стоит алгоритм прямого извлечения данных о рекомбинациях из полностью секвенированных геномов, рассмотренный в главе 1.3.

2. Постановка задачи

Обзор существующих решений показал, что ни один из существующих алгоритмов генетического картирования не является полноценным и удобным способом решения возникающих в современной генетике задач. Тем не менее недостатки алгоритма, предложенного в **1.3** могут быть устранены или сведены к минимуму.

Таким образом, целью данной работы является написание нового алгоритма генетического картирования на базе метода прямого извлечения данных.

Для достижения поставленной цели был сформулирован ряд задач:

1. Доработка алгоритма прямого извлечения данных
2. Верификация полученного алгоритма
 - (a) моделирование и генерация тестовых данных
 - (b) сравнение существующих алгоритмов с полученной версией

3. Доработка алгоритма

3.1. Анализ недостатков и их исправления

3.1.1. Восстановление расположения маркеров по матрице попарных расстояний

В статье [23] утверждается, что алгоритм использует всю информацию о происхождении аллели, которую возможно получить из предоставляемых данных. Для более точного представления процесса работы алгоритма была сформулирована его основная мысль: алгоритм упорядочивает набор маркеров на хромосоме, используя для этого матрицу попарных расстояний. Вообще говоря, непосредственно матрицу попарных расстояний на этапе построения мы получить не можем [30]. Поэтому алгоритм использует её оценку, основанную на частоте рекомбинаций. Как было сказано ранее, частота рекомбинаций между двумя маркерами прямо-пропорциональна расстоянию между ними [31], поэтому матрицу попарных расстояний мы оцениваем матрицей частот рекомбинации.

Так как реализация алгоритма не учитывает кратный кроссинговер, можно предположить, что оценка матрицы попарных расстояний наиболее точна в случае близлежащих маркеров, так как вероятность кроссинговера на отрезке малой длины меньше вероятности кроссинговера на отрезке большей длины. На Рис. 1 показан граф взаимных расстояний с вершинами. С помощью такого представления видно, что по полученным данным можно расположить маркеры в линию.

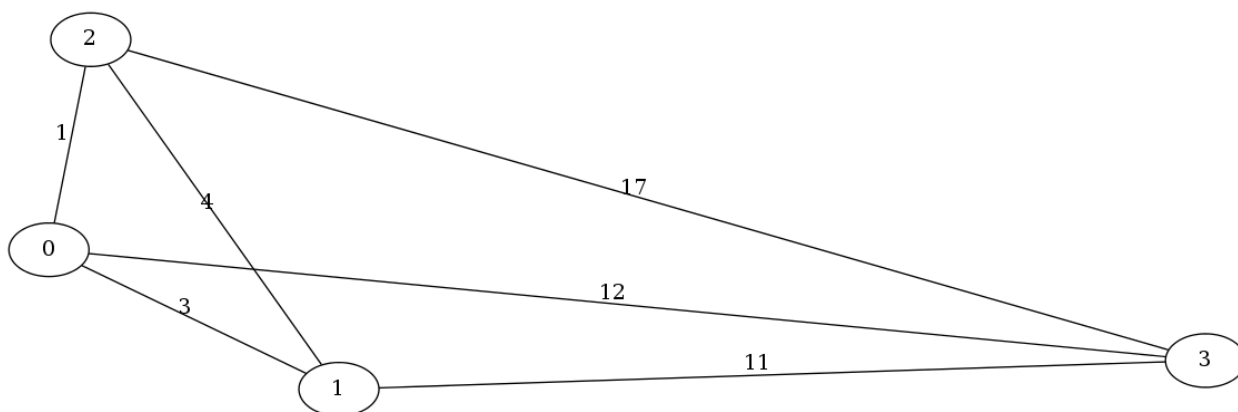


Рис. 1: Попарные расстояния в случае слабо удалённых маркеров

На Рис. 2 показана идентичная ситуация для выборки из четырёх маркеров, далеко расположенных друг от друга.

Таким образом мы показали, что матрица частот рекомбинации является оценкой матрицы попарных расстояний, причём оценка наиболее точна в точках, находящихся близко друг к другу.

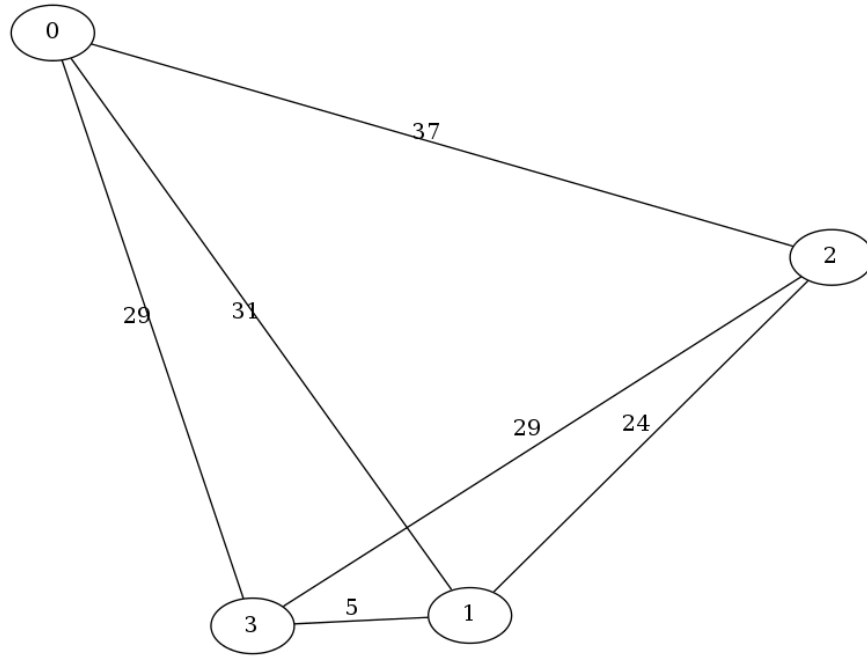


Рис. 2: Попарные расстояния в случае сильно удалённых маркеров

Исходя из приведённого выше замечания, мы считаем, что наиболее правдивыми данными в матрице являются пары маркеров с минимальными значениями частоты рекомбинаций между ними, поэтому начинаем строить карту от центра, представленного двумя самыми близкими маркерами. Более формально этот алгоритм можно записать следующим образом:

Алгоритм 1 Лианеризация маркеров

```

1: function LINMARK
2:    $n \leftarrow \dim(distMatrix)$ 
3:    $(l, r) \leftarrow \operatorname{argmin}_{x,y} distMatrix[x][y]$ 
4:    $order \leftarrow [l, r]$ 
5:   while есть не рассмотренные узлы do
6:      $l \leftarrow first(order)$ 
7:      $r \leftarrow last(order)$ 
8:      $m \leftarrow$  ближайший к  $l$  или  $r$  узел.
9:     if  $m$  ближе к  $r$  then
10:       $order \leftarrow order + [m]$ 
11:    else
12:       $order \leftarrow [m] + order$ 
13:    end if
14:  end while
15:  return order
16: end function

```

— где входным параметром $distMatrix$ является квадратная матрица частот рекомбинаций.

Применяя данный алгоритм, получаем результат в виде порядка маркеров. На Рис. 3 видно, хоть представление и не линейно, что есть некоторые ошибки, и что эти ошибки порядка допустимых транспозиций. Закрашенные узлы — это маркеры, находящиеся не на своих местах. т.е. ошибки. Посмотрев на рисунок можно попробовать

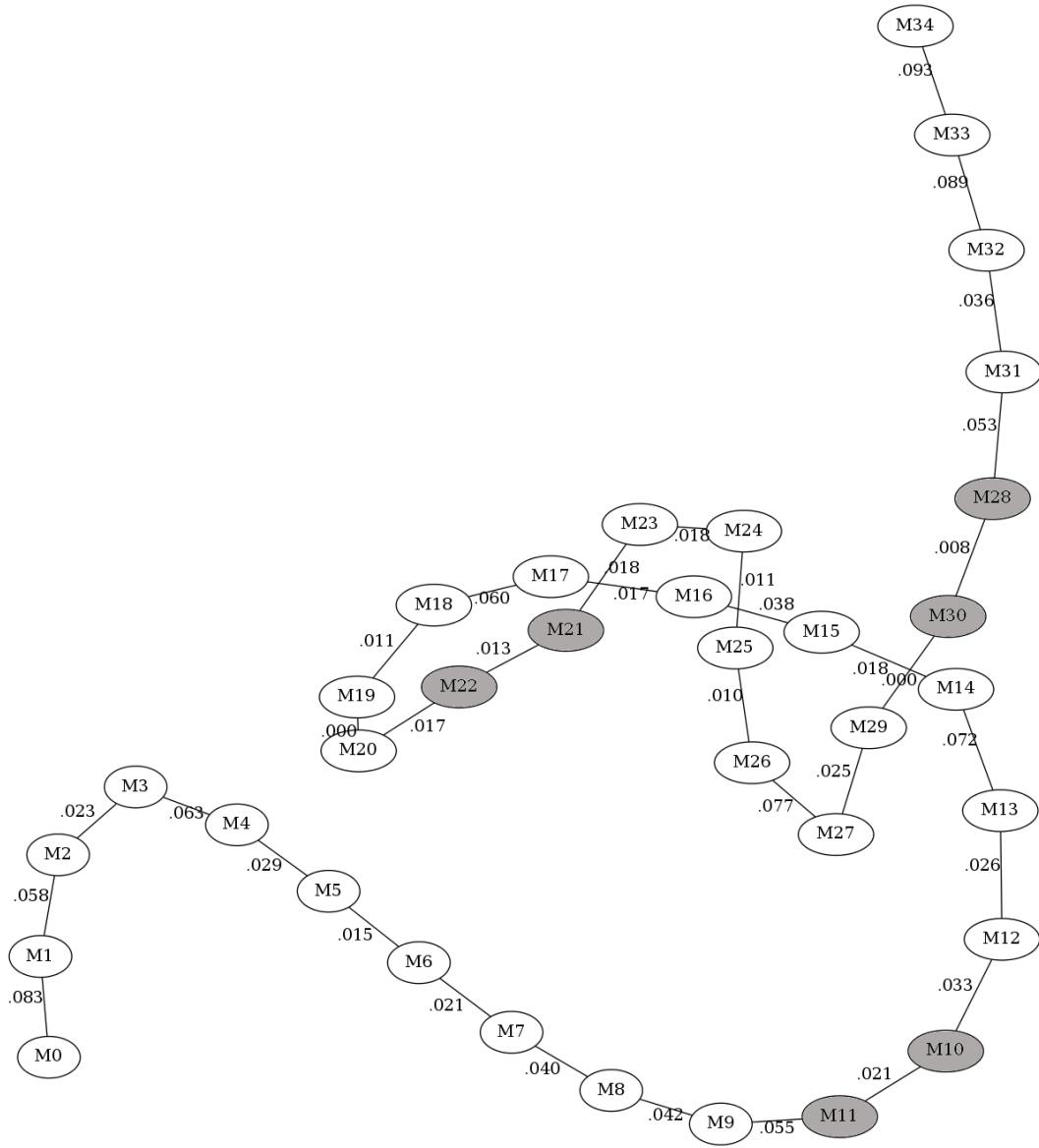


Рис. 3: Порядок маркеров после применения алгоритма вытягивания в линию на семействе из 192 кошек с 35 маркерами

представить изначальную матрицу расстояний в виде графа с вершинами в маркерах и рёбрами с весами, равными частоте рекомбинаций между двумя вершинами (в нашем случае — маркерами). Если вспомнить алгоритм построения минимального остовного дерева Прима [32], который основывается на выборе ребра с минимальным весом, что в нашем случае эквивалентно выбору пары маркеров с минимальным расстоянием между ними среди всех остальных пар, то логичным является попробовать находить порядок маркеров с помощью алгоритма Прима. Заметим так же, что в

случае правильно построенной матрицы попарных расстояний, минимальным основным деревом будет являться искомый порядок, который превратится в линию. На

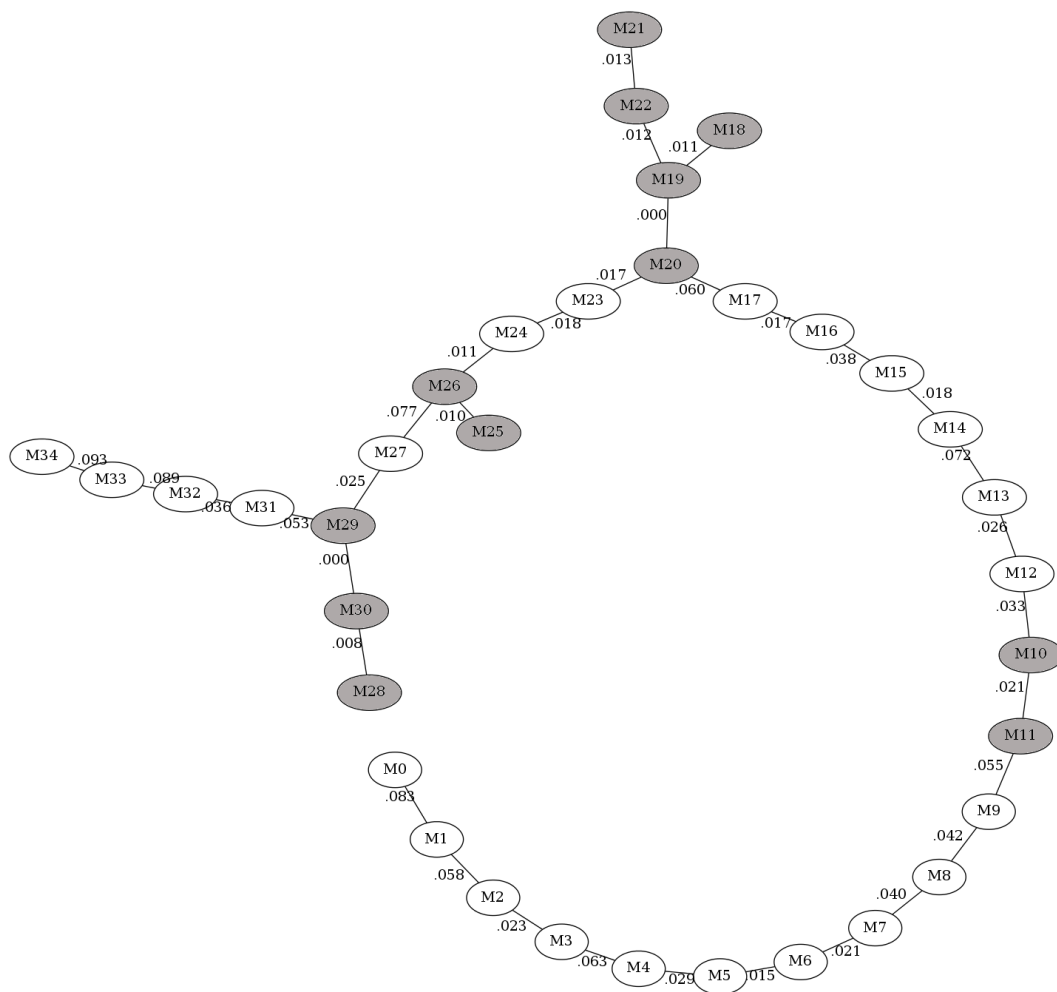


Рис. 4: Минимальное остовное дерево графа попарных расстояний

Рис. 4 представлено построенное с помощью алгоритма Прима минимальное остовное дерево. Не трудно заметить, что те участки, что не имели нарушения порядка в первом варианте остались неизменными, но появились особенности, которые не дают полученному графу называться линией и упорядочить вершины (Аналогично Рис. 3 закрашены те маркеры, порядок которых был определен неверно). Это особенности, вызванные либо парами, между которыми не наблюдалось рекомбинаций (.000 на графе), либо парами маркеров, лежащих друг от друга на большом расстоянии.

Данный пример показал, что в таких задачах применимы основы теории графов и предположения о наибольшем правдоподобии маленьких расстояний. Кроме того, показанный алгоритм при правильной организации хранения графа имеет вычислительную сложность $O(M^2 * \log(M))$, что вносит небольшое улучшение в алгоритм, так как в изначальной версии алгоритма сложность вытягивания матрицы в линию составляла $O(M^3)$. Воспользоваться алгоритмом Прима мы сможем в случае, когда

матрица частот рекомбинаций наиболее точно приближает матрицу попарных расстояний.

3.1.2. Повышение точности матрицы частот рекомбинаций

Рассмотренные выше локальные оптимизации базируются на том, что матрица попарных расстояний, полученная в результате извлечения информации о происхождении аллели правдоподобна. Мы уже показывали, что для далеко отстоящих друг от друга маркеров оценка не верна.

Не трудно заметить, что точность получаемого результата зависит от точности матрицы частот рекомбинаций, так как в нашем случае, именно эта матрица является приближением матрицы попарных расстояний. Как было неоднократно замечено выше, алгоритм, на базе которого мы пытаемся получить новое решение, не учитывает кратность кроссинговера, в связи с чем, матрица частот рекомбинации актуальна только в случае однократного кроссинговера. Резонной является задача извлечения данных о кратных рекомбинациях, так как это явление случается в природе достаточно часто.

Давайте попробуем рассмотреть функцию от расстояния между маркерами $f(d) : \mathbb{R} \rightarrow \mathbb{N}$, возвращающую количество произошедших рекомбинаций. Про эту функцию мы можем смело сказать, что она неубывающая, потому что если рекомбинация произошла на участке, разделяющем два маркера, то на более большом участке рекомбинаций возможно не меньше, чем уже произошли. Однако первоначально нам не известен порядок маркеров и расстояния между ними.

Заметим, что частичные сведения о порядке и расстояниях нам даёт один проход алгоритма по родословной. Естественным образом встаёт вопрос о возможностях уточнения полученного порядка.

Нами было выдвинуто предположение, что базируясь на полученном после одного прохода результате, можно уточнить его рассматривая каждую особь повторно, аккумулируя матрицу частот рекомбинаций, зная примерное расстояние между маркерами.

3.2. Доработанный алгоритм

Так как нахождение матрицы частот рекомбинаций является самым главным этапом алгоритма генетического картирования, а точность полученной матрицы влияет на правильность результата работы всего алгоритма, то основную задачу данной работы можно сформулировать следующим образом: найти алгоритм уточнения матрицы частот рекомбинаций.

Нами было учтено предположение, что можно использовать результат работы изначальной версии алгоритма как предположение о порядке расположения маркеров

на хромосоме. При данном допущении мы можем использовать знания о количестве произошедших рекомбинаций в зависимости от расстояния между ними.

Исходя из этого, получаем дополненный алгоритм, основанный на изначальной версии, входные параметры которого аналогичны с рассмотренным в главе 1.3:

Алгоритм 2 Построение генетических карт с учётом кратности кроссинговера

```

1: function GENMAP2
2:    $order \leftarrow gen\_map(pedegree)$ 
3:    $agg \leftarrow 0_{M,M}$ 
4:   for  $k$  in  $pedegree$  do
5:      $recs \leftarrow 0_{M,M}$ 
6:     for  $i \leftarrow 0; i < M; i++$  do
7:       for  $j \leftarrow i; j < M; j++$  do
8:          $recs[i, j] \leftarrow getRecCount(k, i, j)$ 
9:          $recs[j, i] \leftarrow getRecCount(k, i, j)$ 
10:      end for
11:    end for
12:    for  $i \leftarrow 0; i < M - 1; i++$  do
13:      for  $j \leftarrow 0; j < M - 1; j++$  do
14:        if  $recs[i, j] > recs[i, j + 1]$  and  $recs[i, j + 1] = 0$  then
15:           $recs[i, j + 1] \leftarrow recs[i, j] + 1$ 
16:           $recs[j + 1, i] \leftarrow recs[j, i] + 1$ 
17:        end if
18:        if  $recs[i, j] > recs[i, j + 1]$  and  $recs[i, j + 1] = -1$  then
19:           $recs[i, j + 1] \leftarrow recs[i, j]$ 
20:           $recs[j + 1, i] \leftarrow recs[j, i]$ 
21:        end if
22:      end for
23:    end for
24:     $agg \leftarrow agg + recs$ 
25:  end for
26:  return Linmark(agg)
27: end function

```

Новый алгоритм реализует описанную выше схему. Покажем, что эта схема оптимальная и более быстрая, чем наиболее популярные сейчас алгоритмы. А так же обоснуем преимущества данной версии алгоритма по сравнению с предыдущей реализацией.

4. Сравнение алгоритмов

Существует два подхода к сравнению биологических алгоритмов:

- сравнение с использованием реальных данных.
- сравнение с использованием синтетических данных.

Зачастую провести корректное сравнение при помощи реальных данных не удаётся по следующим причинам:

- Биологических данных не много.

Секвенирование молекул ДНК не производится для некоторых видов по причине дороговизны или отсутствия научного интереса к этим данным.

- Данные из разных источников могут быть не сопоставимы.

Например, достаточно мало данных, которые отличаются минимальным набором параметров. Трудно найти секвенированные родословные 250 собак, если в одном случае нужны 250 особей только белого цвета, а в другой 250 собак, отличающихся цветом глаз. Чаще всего встречаются разрозненные наборы данных, с разными количеством и видами маркеров, количеством поколений и т.п.

- Для биологических данных характерны ошибки.

Хоть и утверждается, что вероятность ошибок при секвенировании минимальна, они все-таки случаются. Получение природного материала не может быть однозначным по определению, так как велик человеческий фактор и недостатки биологического программного обеспечения.

- Получение биологических данных дорого.

Секвенирование особей и тем более целых родословных стоит дорого. Например, полное секвенирование генома человека стоит не менее 1000 USD [33]. Для целей тестирования и отладки методов это несоизмеримые затраты.

Синтетические данные лишены описанных выше недостатков, но при неправильной модели синтеза данных они могут не соответствовать действительности, исказить или вовсе потерять биологический смысл.

Именно поэтому для сравнения алгоритмов между собой, а так же для верификации новой версии алгоритма нами была выбрана следующая стратегия:

Этап 1: Сравнить на синтетических данных, чтобы обнаружить сильные и слабые стороны алгоритма

Этап 2: Верифицировать на реальных данных с целью проверить работоспособность

Сравнение существующих алгоритмов на синтетических данных является важной задачей, так как это позволяет выявить немало полезной информации, позволяющее выявить вероятные ошибки, проверить и отладить алгоритм. Кроме того, данные можно генерировать для акцентирования внимания на исследуемой проблеме. Например, в случае исследования поведения программы на данных, полученных путём моделирования кратного кроссинговера.

4.1. Генерация тестовых данных

Генерация тестовых данных в случае генетики является задачей, требующей аккуратности, так как во время моделирования сгенерированные тестовые данные могут потерять биологический смысл. Генерируя тестовое множество особей, нам необходимо учитывать эти обстоятельства. В Алгоритме 3 показан алгоритм генерации родословных.

4.2. Процесс сравнения и анализа

Для наиболее полного сравнения алгоритмов стоит выделить ряд биологических аспектов, влияющих на результаты работы алгоритмов:

- наличие большого количества неинформативных пар (много гомозиготных локусов)
- наличие неоднократного кроссинговера
- его чётность

Не стоит забывать, что также немаловажным для скорости выполнения алгоритма является порядок количества маркеров и мощность родословной.

В связи с этим рассмотрим сочетания биологических особенностей, влияющих на качество результатов алгоритмов, с объёмом данных. В Таблице 1 на пересечении строки и столбца находится имя тестового множества, основанного на родословной, содержащей столько особей, сколько находится в названии строки и показывающем столько макреков, сколько в названии столбца.

организмы маркеры	30	40	50
200	o200m30	o200m40	o200m50
300	o300m30	o300m40	o300m50
400	o400m30	o400m40	o400m50

Таблица 1: Входные данные

Стоит заметить так же, что помимо количественных характеристик, родословные имеют такие параметры, как количество кроссинговеров, вероятность кроссинговеров

Алгоритм 3 Генерация родословной

```
1: procedure GENPEDIGREE(n_gen, crossover_strategy)
2:   population  $\leftarrow$  []
3:   founders  $\leftarrow$  список прародителей
4:   population  $\leftarrow$  population + founders
5:   for gen  $\leftarrow$  0; gen < n_gen; gen ++ do
6:     (males, females)  $\leftarrow$  divide(population)
7:     mother  $\leftarrow$  choice(females)
8:     father  $\leftarrow$  choice(males)
9:     child  $\leftarrow$  CrossBreed(mother, father)
10:    добавить child к population
11:   end for
12:   print population
13: end procedure
14:
15: function CROSSBREED(mother, father)
16:   mothergamet  $\leftarrow$  Crossover(motherchromosomes, crossover_strategy)
17:   fathergamet  $\leftarrow$  Crossover(fatherchromosomes, crossover_strategy)
18: end function
19:
20: function CROSSOVER(chromosomeA, chromosomeB, crossover_strategy)
21:   crossovers  $\leftarrow$  точки кроссинговера в соответствии с crossover_strategy
22:   gametA  $\leftarrow$  []
23:   gametB  $\leftarrow$  []
24:   for position in chromosome do
25:     if position in crossovers then
26:       swap chromosomeA, chromosomeB
27:       gametA  $\leftarrow$  gametA + chromosomeA[position]
28:       gametB  $\leftarrow$  gametB + chromosomeB[position]
29:     end if
30:   end for
31:   return random(gametA, gametB)
32: end function
```

и т.п. Следующие тестовые множества были сгенерированы с целью показать точность того или иного алгоритма.

Тестовые множества для проверки точности построения с учетом кратного кроссинговера:

nocross — 200 кошек, 30 маркеров, нет кроссинговеров

singlecross — 200 кошек, 30 маркеров, только однократные кроссинговеры

manycross — 200 кошек, 30 маркеров, кратные (больше одного) кроссинговеры кроссинговеров

cats — 192 кошки, 35 маркеров — реальные данные.

При верификации алгоритма мы будем обращать внимание на два параметра: скорость решения задачи и точность полученных результатов. Т.о. более точно процесс сравнения алгоритмов можно описать следующим образом:

Шаг 1: Для каждого тестового множества запустить поочередно исследуемые реализации алгоритмов

Шаг 2: Засечь время выполнения задачи и затем записать их в соответствующую таблицу.

Шаг 3: Проверить количество ошибок в полученном ответе и записать их в соответствующую таблицу

4.3. Обобщённые данные

Для осуществления сравнения существующих алгоритмом генетического картирования нами были выбраны наиболее популярные программные пакеты:

- CRIMAP: реализует алгоритм Ландера-Грина
- FASTLINK: реализует алгоритм Элстона-Стюарта
- genmap: реализация предложенного в данной работе алгоритма

Используя описанную главой выше схему, получает следующие данные о скорости, демонстрируемые в Таблице 2. Не трудно заметить заметный выигрыш нашего алгоритма по сравнению с остальными.

В Таблице 3 мы использовали другое тестовое множество, чтобы показать, как кратность кроссинговера влияет на правильность получаемых результатов. Из полученных данных видно, что ошибка, получаемая после результата работы нашего

	genmap.py	CRIMAP	FASTLINK
o200m30	1c	10m	1h
o200m40	1c	10m	2d
o200m50	2c	10m	∞
o300m30	2c	1d	2h
o300m40	3c	1d	2d
o300m50	3c	1d	∞
o400m30	5c	∞	3h
o400m40	6c	∞	3d
o400m50	8c	∞	∞

Таблица 2: Скорость работы алгоритмов

	genmap	CRIMAP	FASTLINK
nocross	0	0	0
singlecross	0	0	0
manycross	4	2	3
cats	1	0	0

Таблица 3: Количество ошибок

алгоритма того же порядка, что и у остальных. Для сравнения, в Таблице 4 показаны результаты работы предыдущей версии алгоритма, что наглядно показывает, что предлагаемые в этой работе изменения действительно являются улучшениями.

Запуск на этих данных показал, что алгоритм предложенный в данной работе работает намного быстрее других, что его качество уменьшается с увеличением эффекта от кратных кроссинговеров, но для реальных биоданных его точности более чем хватает.

	old genmap	CRIMAP	FASTLINK
nocross	0	0	0
singlecross	1	0	0
manycross	14	2	3
cats	1	0	0

Таблица 4: Количество ошибок в старой версии genmap

Заключение

Результаты

В ходе выполнения дипломной работы нами был предложен доработанный алгоритм прямого извлечения информации о рекомбинациях из секвенированного генома, позволяющий получать генетические карты в случае одно- и многократного кроссинговера. Для верификации нового алгоритма был реализован механизм тестирования методов построения генетических карт, с помощью которого были наглядно представлены преимущества нашей реализации алгоритма.

Актуальность полученных результатов

Применение генетических карт получило достаточно широкое распространение в сфере диагностирования наследственных заболеваний. Исследование механизмов наследования, а так же выявление его закономерностей позволяет получить информацию о предрасположенности у человека (и не только) к тем или иным отклонениям, которые трудно выявить до начала проявления симптомов. Получение информации о генной предрасположенности до начала заболевания позволяет модифицировать лечение и образ жизни, избегая или купируя их дальнейшие проявления.

Кроме того, генетические карты применяются в исследованиях процессов эволюции. Рассматривая генетические карты двух достаточно близких видов, можно извлечь информацию о том, в каком направлении эволюционировала жизнь и какие гены передавались в первую очередь.

Из-за недостатков современных методов секвенирования, отсутствия возможности секвенировать человека большими поколениями и дороговизны этого процесса, единственными способами построить генетическую карту являются рассмотренные выше алгоритмы. Предложенный нами алгоритм позволяет строить карты быстрее, чем альтернативные алгоритмы, и точнее, чем его предшественник.

Список литературы

- [1] Б.В. Петровский. Краткая медицинская энциклопедия. *Сов. Энциклопедия*, 1989.
- [2] Anthony JF Griffiths. *An introduction to genetic analysis*. Macmillan, 2005.
- [3] Anthony JF Griffiths, Jeffrey H Miller, David T Suzuki, and Richard C Lewontin. *An introduction to genetic analysis*. 2000.
- [4] Б Албертс, Д Брей, Дж Льюис, К Роберте, and Дж Уотсон. Молекулярная биология клетки. В трех томах, 1994.
- [5] Strachan T. and Read E. *Human Molecular Genetics*. Garland Science, New York, 4 edition, 2009.
- [6] Thomas Hunt Morgan, Alfred Henry Sturtevant, Hermann Joseph Muller, and Calvin Blackman Bridges. *The mechanism of Mendelian heredity*. Holt, New York, 1922.
- [7] Jon Schiller. *Genome Mapping: To Determine Disease Susceptibility*. CreateSpace, 2010.
- [8] Harriet B Creighton and Barbara McClintock. A correlation of cytological and genetical crossing-over in *zea mays*. *Proceedings of the National Academy of Sciences of the United States of America*, 17(8):492, 1931.
- [9] Susan Goldhor. Genetics and evolution—selected papers of ah sturtevant. *The Yale journal of biology and medicine*, 34(5):528, 1962.
- [10] Alfred Henry Sturtevant, George Wells Beadle, et al. An introduction to genetics. *An introduction to genetics.*, 1939.
- [11] Daniel Hartl and Maryellen Ruvolo. *Genetics*. Jones & Bartlett Publishers, 2011.
- [12] George M Malacinski. *Essentials of molecular biology*. Jones & Bartlett Learning, 2005.
- [13] В. В. Хвостова and Ю. Ф. Богданов. *Цитология и генетика мейоза*. Наука, Москва, 1975.
- [14] MARY SARA McPEEK. An introduction to recombination and linkage analysis. In *Genetic Mapping and DNA Sequencing*, pages 1–14. Springer, 1996.
- [15] Gerard D Schellenberg, Thomas D Bird, Ellen M Wijsman, Harry T Orr, Leojean Anderson, Ellen Nemens, June A White, Lori Bonnycastle, James L Weber, M Elisa Alonso, et al. Genetic linkage evidence for a familial alzheimer’s disease locus on chromosome 14. *Science*, 258(5082):668–671, 1992.

- [16] Isabelle Oberle, Giovanna Camerino, Roland Heilig, Lelia Grunebaum, Jean-Pierre Cazenave, Calogero Crapanzano, Pier M Mannucci, and Jean-Louis Mandel. Genetic screening for hemophilia a (classic hemophilia) with a polymorphic dna probe. *New England Journal of Medicine*, 312(11):682–686, 1985.
- [17] MARY ANNE ANDERSON, RUDOLPH E TANZI, PAUL C WATKINS, KATHLEEN OTTINA, MARGARET R WALLACE, ALAN Y SAKAGUCHI, ANNE B YOUNGH, IRA SHOULSONH, and ERNESTO BONILLAH. *A polymorphic DNA marker genetically linked to Huntington's disease*, volume 51. Oxford University Press, 2004.
- [18] Eric S Lander, Philip Green, Jeff Abrahamson, Aaron Barlow, Mark J Daly, Stephen E Lincoln, and Lee Newburg. Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, 1(2):174–181, 1987.
- [19] Phil Green, Kathy Falls, and Steve Crooks. *Documentation for CRI-MAP*, 2.4 edition, 1990.
- [20] Fastlink package.
- [21] Leonid Kruglyak, Mark J Daly, Mary Pat Reeve-Daly, and Eric S Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American journal of human genetics*, 58(6):1347, 1996.
- [22] Maáyan Fishelson and Dan Geiger. Exact genetic linkage computations for general pedigrees. *Bioinformatics*, 18(suppl 1):S189–S198, 2002.
- [23] Сысоев С.С. *Эффективный алгоритм построение генетических карт по полнотью секвенированным участкам геномов*. Издательство С.-Петербургского Университета, СПб, 2013.
- [24] *Биологический энциклопедический словарь*. Советская Энциклопедия, Москва, 1986.
- [25] Robert C Elston and John Stewart. A general model for the genetic analysis of pedigree data. *Human heredity*, 21(6):523–542, 1971.
- [26] Eric S Lander and Philip Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, 84(8):2363–2367, 1987.
- [27] E.S. Lander and P. Green. *The Progress of Genetic Research*, volume 84 (8). National Academy of Sciences, USA, 1987.

- [28] Jeffrey R O’Connell. Rapid multipoint linkage analysis via inheritance vectors in the elston-stewart algorithm. *Human heredity*, 51(4):226–240, 2001.
- [29] Maáyan Fishelson and Dan Geiger. Optimizing exact genetic linkage computations. *Journal of Computational Biology*, 11(2-3):263–275, 2004.
- [30] AJ Bohonak. Ibd (isolation by distance): a program for analyses of isolation by distance. *Journal of Heredity*, 93(2):153–154, 2002.
- [31] Piet Stam. Construction of integrated genetic linkage maps by means of a new computer package: Join map. *The Plant Journal*, 3(5):739–744, 1993.
- [32] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, Clifford Stein, et al. *Introduction to algorithms*, volume 2. MIT press Cambridge, 2001.
- [33] Martin Kircher and Janet Kelso. High-throughput dna sequencing—concepts and limitations. *Bioessays*, 32(6):524–536, 2010.