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

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]. 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 [3]. К сожалению, современные методы секвенирования не в состоянии предоставить информацию о полной нуклеатидной последовательности в рамках конкретной хромосомы [4]. Сама по себе нуклеатидная последовательность не содержит прямой информации о происхождении того или иного однозначно идентифицируемого участка (маркера), и становится затруднительным определить, от какого предка был унаследован тот или иной признак. Для определения шаблонов наследования и выявления его принципов предназначены генетические карты хромосом[5].

Генетическая карта - это схема или порядок расположения маркеров на хромосоме (здесь и далее подразумевается структурный маркер, т.е. маркер, имеющий отличимое нуклеатидное представление, которое позволяет его идентифицировать) и генов. Зачастую, наличие генов не так важно, как наличие маркеров, потому что гены имеют свойство наследоваться не полностью [6]. Идея создать генные карты принадлежит Томасу Моргану, внёсшему неоспоримый вклад в теорию наследственности. Идея была основана на явлении сцепленного наследования генов. Из-за мейотического кроссинговера, который делает невозможным полностью скоррелированное наседование и влияет на расхождение сцепленных генов по разным гаметам, появилось предположение о связи физического расстояния между маркерами и их взаимодействии при наследовании. Это предположение оправдалось, и в 1913 году ученик Моргана Альфред Стёртевант построил первую генетическую карту на основе данных Drosophila melanogaster [7, 8].

Кроссинговер на этапе мейоза вносит возмущение в сцепленное наследование [9], 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. Генетическое расстояние не трудно перевести в физическое, но чаще всего в этом нет необходимости, так как для механизма наследования существенен порядок [10]. Единицей измерения расстояния является 1 сантиморган. Стоит упомянуть, что задача построения генетической карты хромосомы имеет смысл для диплоидных особей и лучше решается, когда исследуемые особи находятся в родстве. Задача так же решается проще

на семействах особей, которые размножаются в большой скоростью, поэтому большее количество карт на данный момент имеют хромосомы дрозофил, кошек, мелких грызунов и насекомых. Нас же интересуют генетические карты человека.

Современные средства генетического картирования позволяют построить карту по форматированному файлу родословной. Причём не обязательно иметь полную информацию о степени родства, половой принадлежности и возрасте. Более того, секвенирование допускает ошибки в порядке генов.

Общепризнанными и наиболее распространёнными методами генетического картирования являются:

- Алгоритм Элстона-Стюарта
- Алгоритм Ландера-Грина

Программные средства, основанные на вышеуказанных алгоритмах хорошо решают поставленную задачу для взятом у человеческих особей материале при сравнительно небольшом (по сравнению с известными науке видами маркеров) количеством маркеров [Алгоритм Элстона-Стюарта] и небольшом размере исследуемой семьи [Алгоритм Ландера-Грина]. Практические требования современной медицины приводят к необходимости строить генетические карты по все большим и большим наборам маркеров. Вычислительная сложность алгоритма Элстона-Стюарта растёт экспоненциально по этому параметру, в результате чего генетическое картирование становится несущественным на практике. Алгоритм Ландера-Грина позволяет исследовать большее количество маркеров, но на меньших семействах, так как время работы этого алгоритма экспоненциально растёт с ростом количества наблюдаемых в родословной особей.

В 2013 году в работе [11] был предложен алгоритм прямого извлечения информации о генетических расстояниях между маркерами и построения генетической карты без учёта кратности кроссинговера. Вычислительная сложность данного алгоритма оценивается степенной функцией как от количества маркеров, так и от мощности родословной, что делает его перспективнее ранее упомянутых алгоритмов для нужд современной генетики. При этом алгоритм прямого извлечения данных имеет ряд существенных недостатков, а именно:

1. Плохая теоретическая обоснованность

   В работе [11] утверждается, что алгоритм извлекает всю возможную информацию о рекомбинациях в исследуемых особях, в главе [NUMBER] мы покажем, что это утверждение вообще говоря не верно.

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

   что существенно повышает ошибочность алгоритма, так как в природе неоднократный кроссинговер — явление, встречающееся достаточно часто

Исходя из этого, цель этой работы можно сформулировать следующим образом: предложить свой алгоритм построения генетических карт на основе алгоритма [11],

- усовершенствовав алгоритм в части учёта кратных рекомбинаций
- верифицируя полученный алгоритм на реальных и искусственно синтетических данных
- сравнив новый алгоритм с предшественниками

1. Обзор существующих решений

Основные понятия

Для того, чтобы рассматривать существующие подходы, нужно более формально поставить задачу, которую они решают. Нам потребуются следующие понятия [12]:

- *маркер* или *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 которого представлена одинаковыми аллелями
- *гетерозигота* — диплоидная особь, копия генов которого представлена разными аллелями.

  Поясню введённые выше термины 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 позиции 1, так как имеет разные маркеры от отца и матери, а если бы в позиции 2 находилась строка **BB**, по 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.

  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 гаметы, которая передастся 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

Используя эти определения перейдём к алгоритмическому смыслу задачи и рассмотреть её решения. Имея на входе данные о родословной, количестве исследуемых маркеров, нуклеотидных последовательностях всех особей, а так же их родственной связи, генетическую карту можно построить следующими алгоритмами, имеющими программную реализацию:

- Алгоритм Ландера — Грина [13]
- Алгоритм Элстона — Стюарта [14]
- Построение генетических карт по полностью секвенированным участкам геномов [11]

1.1. Алгоритм Элстона-Стюарта

Описание - достоинства - недостатки

1.2. Алгоритм Ландера-Грина

Описание - достоинства - недостатки

1.3. Построение генетических карт по полностью секвенированным участкам геномов

Описание из статьи - достоинства - недостатки

Построение генетических карт

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

2. Постановка задачи

Обзор существующих решений показал, что ни один из существующих алгоритмов генетического картирования не является полноценным и удобным для решения возникающих в современной генетике задач. Тем не менее недостатки алгоритма, предложенного в 1.3 могут быть усранены 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 полученный алгоритм
   (a) реализовать возможность моделирования и генерации тестовых данных
   (b) сравнить существующие алгоритмы с новой версией

3. Доработка алгоритма

3.1. Анализ недостатков и гипотезы их исправления

Для того, чтобы добиться улучшений алгоритма был произведён его тщательный анализ. Помимо того, что логика алгоритма не подразумевает возможность кратного кроссинговера, было выявлено проявление этого надочета в получаемых данных.

картинка с большими расстониями, а как раз на больших и могло сбиться все

Кроме того, если матрице попарных расстояний представить в виде графа и воспользоваться предположением, что в случае нахождения всех вершин на одной прямой, результатом будет являться минимальное остовное дерево, то проверив, получаем, что результат совпадает с полученным ранее методом “вытягивания матрицу в линию”.

Картинка с вытягиванием.

Картинка с остовом.

Так же зная, что количество рекомбинаций есть неубывающая функция от расстояния между двумя маркерами, можно сделать вывод, что случай неинформатиынх мейозов можно покрыть.

тут наша с вами схема с матрицами и всем таки с картинками, чтобы понятно было

3.2. Доработанный алгоритм

Новый алгоритм: Псевдокод

4. Сравнение алгоритмов

Сравнение существующих алгоритмов на синтетических данных является важной задачей, так как это позволяет выявить немало полезной информации, такой как наиболее подходящие случаи для использования, вероятные ошибки, вычислительную сложность. Кроме того, данные можно генерировать дл акцентирования внимания на исследуемой проблеме. Например, в случае исследования поведения программы на данных, полученных путём моделирования кратного кроссинговера. Заметим, что подобную информацию трудно получить из естественных результатах секвенирования. Кроме того, параметризация генерируемых данных упрощает проверку результата работы алгоритма.

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. При решении этой задачи нами было выявлено 2 принципиально разных подхода:

- генерирование данных по заданным параметрам
- генерирование данных на основе известного результата

В первом случае, генерируя тестовое множество особей, нам необходимо понимать, каким образом выглядит итоговая и желаемая генетическая карта.Во втором, имея карту, легко сравнивать получаемые результаты работы алгоритма. Этот подход заведомо сложнее, так как никто не гарантирует биекцию между множествами входных данных и множеством результатов работы алгоритма на этих входных данных. В связи с этим мы выбрали первый подход.

Входные параметры:

- количество особей-прародителей (N)
- количество прямых потомков (детей) (F1)
- количество поколений (gen_count)
- итоговое количество особей в родословной (common_N)
- вероятность рекомбинации при мейозе (rec_prob)
- количество рассматриваемых маркеров (markers_count)

Особь в нашем случае, это объект класса Organism
// тут будет код и его описание в виде листинга
Алгоритм генерации родословной:

1. Создаём пустой список pedigree длины common_N
2. Генерируем m идентификаторов для маркеров
3. На основании количества особей-прародителей N генерируем аллели с m маркерами и задаём особям уникальный идентификатор
4. $N/2$ особям указываем мужской пол, а оставшимся женский
5. Записываем в массивы possible_fathers и possible_mothers мужские и женские особи соответственно
6. Пополняем pedigree получившимися N особями
7. Пока $F1 > 0$:
   (a) mother <- possible_mothers[random() mod $N/2$]
   (b) father <- possible_fathers[random() mod $N/2$]
   АЛИНА, ПРИДУМАЙ ЧИТАЕМЫЙ ПСЕВДОКОД, ЭТОТ ДАЖЕ СЕЙЧАС НЕ ОЧЕНЬ, А ТЫ ЕЩЕ ДО ОБРАЗОВАНИЯ ГАМЕТ НЕ ДОШЛА

4.2. Процесс сравнения и анализа

Для наиболее хорошего и 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 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.

ТАБЛИЧКА

4.3. Обобщённые данные

ЕЩЁ ОДНА ТАБЛИЧКА

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 моделировать лечения 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 фауна и какие гены передавались в первую очередь.

Из-за недостатков современных методов секвенирования, а так же из-за отсутствия возможности секвенировать гаметы, 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] *Краткая Медицинская Энциклопедия.* Советская Энциклопедия, Москва, 1989.

[2] Suzuki D.T. Griffiths A.J.F., Miller J.H. *An Introduction to Genetic Analysis.* W. H. Freeman, New York, 7 edition, 2000.

[3] Альбертс Б., Брей Д., Льюис Дж., Рэфф М., Робертс К., and Уотсон Дж. *Молекулярная биология клетки: в трех томах.*, volume 1. Мир, Москва, 2 edition, 1994.

[4] Strachan T. and Read E. *Human Molecular Genetics.* Garland Science, New York, 4 edition, 2009.

[5] T.H. Morgan, A.H. Sturtevant, H.J. Muller, and C.B. Bridges. *The Mechanism of Mendelian Heredity.* Henry Holt, New York, 1923.

[6] W Bateson. The progress of genetic research. In *Report of the Third 1906 International Conference on Genetics: Hybridization (the cross-breeding of genera or species), the cross-breeding of varieties, and general plant breeding.*, London, 1907. Royal Horticultural Society.

[7] E.B Lewis, editor. *Genetics and Evolution: Selected Papers of A.H. Sturtevant.* Freeman and Company, San Francisco, 1961.

[8] A.H. Sturtevant and Beadle. G.W., editors. *An Introduction to Genetics.* Saunders Company, Philadelphia, PA, 1940.

[9] McClintock B., editor. *A Cytological Demonstration Of The Location Of An Inherchange Between Two Non-Homologous Chromosomes Of Zea Mays.* 1930.

[10] Daniel L. Hartl and Maryellen Ruvolo, editors. *Genetics: Analysis of Genetics and Genomes.* Jones & Bartlett, Burlington, 2012.

[11] Сысоев С.С. Эффективный алгоритм построение генетических карт по полностью секвенированным участкам геномов. In *Report of the Third 1906 International Conference on Genetics: Hybridization (the cross-breeding of genera or species), the cross-breeding of varieties, and general plant breeding.*, volume 84 (8), page 2363–2367, СПб, 2013. National Academy of Sciences.

[12] М.С. Гиляров., editor. *Биологический энциклопедический словарь.* Советская Энциклопедия, Москва, 1986.

[13] E.S. Lander and P. Green. The progress of genetic research. volume 84 (8), page 2363–2367, USA, 1987. National Academy of Sciences.

[14] R. C. Elston and J. Stewart. *A general model for the genetic analysis of pedigree data.* Hum Hered, 1971.