Санкт-Петербургский Государственный Университет
Математико-механический факультет

Кафедра системного программирования

Программные инструменты и алгоритмы для определения информации о последовательности нуклеотидов по ее местоположению в геноме человека.

Курсовая работа студента 445 группы
Алеева Алексея Валерьевича

Научный руководитель ……. Н.И. Вяххи
ассистент кафедры информатики

Санкт-Петербург
2011

Содержание

Введение.....3
Постановка задачи.....5
Обзор существующих решений.....6
Ensembl.....6
UCSC Table Browser.....9
UCSC Genome Browser.....11
Итоги обзора существующих решений.....13
Реализация собственного решения
поставленной задачи.....14
Заключение.....19
Используемые источники.....20

Введение

За последние годы биология начала превращаться в науку, «богатую данными». Размер бактериального генома — миллионы нуклеотидов, высшего животного — сотни миллионов или миллиарды. Транскриптомика, изучающая активность генов, получает данные о концентрациях десятков тысяч матричных РНК, протеомика — о сотнях тысяч пептидов и белок-белковых взаимодействиях. С таким количеством информации нельзя работать вручную, поэтому обработку информации необходимо автоматизировать.

Исходная информация рождается в тысячах экспериментов, независимо проводимых по всему миру с разными целями. Полученные данные экспериментатор сравнивает со всем набором уже имеющихся в банке (одним из таких банков является крупнейший банк данных генных последовательностей EMBL Nucleotide Databank [1]). Например, определив в опыте нуклеотидную последовательность нового гена с неизвестной функцией, исследователь может узнать, имеются ли у других организмов похожие гены, и какую функцию они выполняют. Может оказаться, что ген человека, связанный с каким-либо заболеванием, похож, например, на бактериальный ген с известной функцией. Это даст нить для дальнейших исследований по определению роли гена в заболевании. Если исследователь не нашел в банке открытой им последовательности, он посылает ее туда, пополняя банк.

Существует множество приложений, обеспечивающих доступ к геномной информации и работе с ней, но зачастую они недостаточно удобны, либо предоставляют недостаточно широкую функциональность.

Проект Genome Query создан для того, чтобы помочь ученым или просто интересующимся данной тематикой людям в решении каких-либо поисковых задач, проверке некоторых статистических гипотез или нахождении информации о геноме человека, обеспечивая удобный доступ и работу с геномной информацией, при этом предоставляя широкую функциональность.

В настоящее время в проекте поддерживается поиск заданной нуклеотидной последовательности в геноме человека, как точный, так и неточный, а также

предоставляется аннотационная информация о различных элементах генома человека, в частности, о генах.

Проект предоставляет веб-сервис для получения вышеуказанной информации, а также web-API, ведётся разработка библиотеки реализованных алгоритмов для поиска последовательностей.

Одной из основных областей биоинформатики является аннотация геномов. В контексте геномики аннотация — процесс маркировки генов и других объектов в последовательности ДНК. К аннотации генома относят описание функциональных и структурных характеристик генома, местонахождение кодирующих участков генов в геноме, регуляторных элементов, регулирующих транскрипцию и другие функции генома, особенностей функционирования генома, в частности, тканеспецифичности экспрессии генов, так называемых профилей их экспрессии, взаимосвязей между генами и других функциональных свойств генома.

Мы рассмотрим частный случай данной задачи, а именно – получение информации о последовательности нуклеотидов по ее местоположению в геноме человека. Исследователя может интересовать, какие гены находятся в хромосоме или ее части. Другим примером является следующая ситуация: пациента прокариотипировали (установили его кариотип, то есть количество и морфологию хромосом) и увидели, что у него потеряна часть короткого плеча, скажем, 18-й хромосомы. Указав примерный диапазон потерянного участка, хотелось бы узнать, чем такая потеря грозит.

Постановка задачи

Задача состоит в том, чтобы провести анализ существующих решений задачи о получении информации о последовательности нуклеотидов по ее местоположению в геноме человека и, по результатам обзора, принять решение о добавлении данной функциональности в рамках проекта Genome Query.

Обзор существующих решений

Для обзора были выбраны наиболее известные сервисы, позволяющие получить информацию об участке генома человека по его позиции. Также существует ряд сервисов, использующих те же базы данных, что и рассмотренные, поэтому они не были включены в обзор.

Ensembl

Ensembl [2] является совместным научным проектом Европейского института биоинформатики (EBI) и Wellcome Trust Sanger Institute, который был запущен в 1999 году в ответ на предстоящее окончание проекта «Геном человека» [3]. После 10 лет существования, целью Ensembl остается обеспечить централизованный ресурс для генетиков, молекулярных биологов и других исследователей, изучающих геномы нашего собственного вида и других позвоночных организмов. Ensembl является одним из наиболее известных браузеров для поиска геномной информации.

Пользователь может указать тип генома, хромосому и диапазон в ней. В результате будет показана информация об элементах генома в данной его части, а также вокруг нее.

На рис.1 показано поле ввода для указания интересующего нас участка, а на рис.2 – результат работы сервиса.

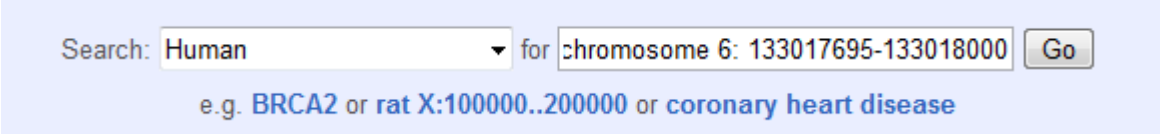


Рис.1 Поле ввода в Ensembl

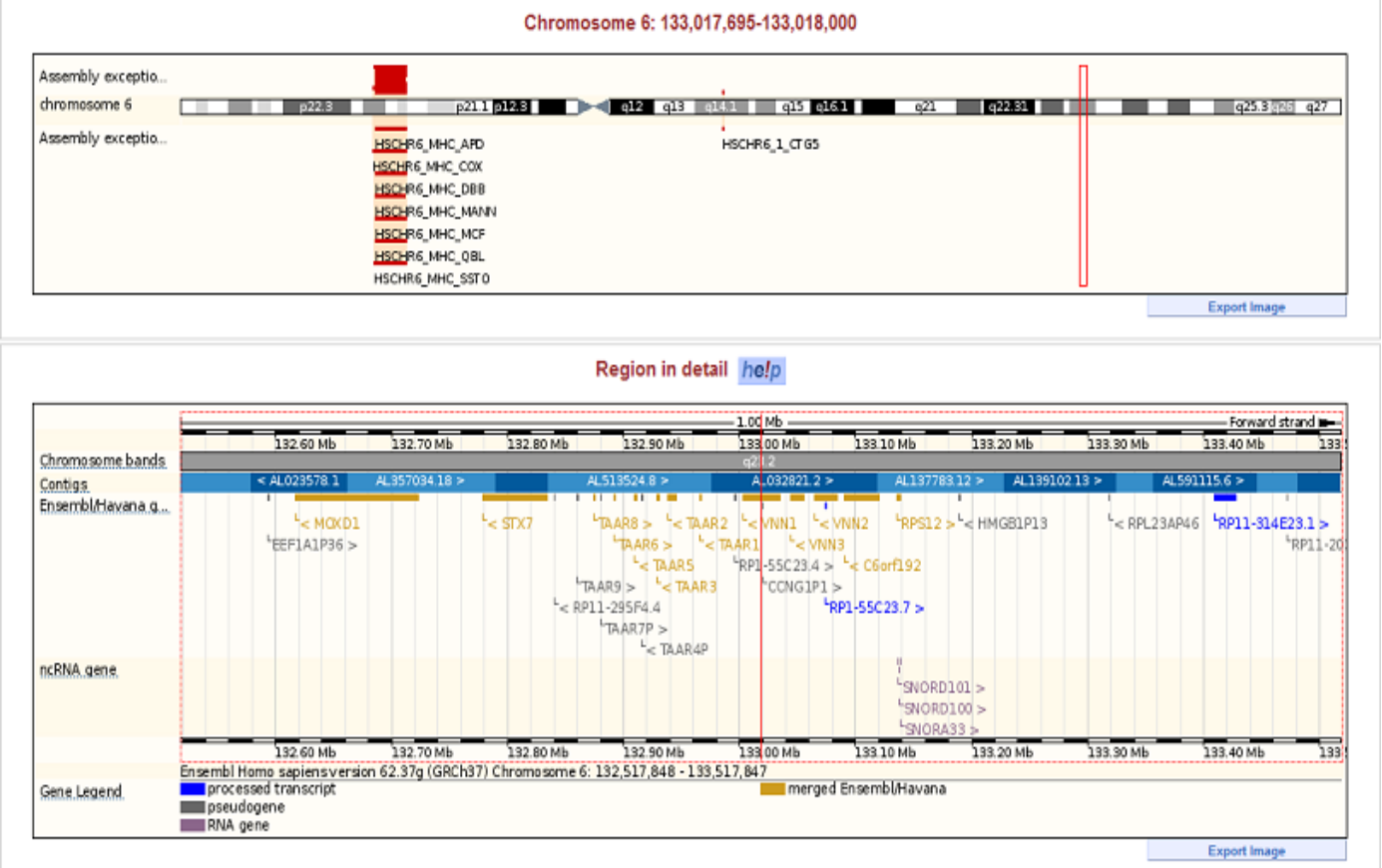


Рис.2 Результат работы

Щелкнув мышью по интересующим нас элементам, будет показана некоторая информация о них, это показано на рис.3:

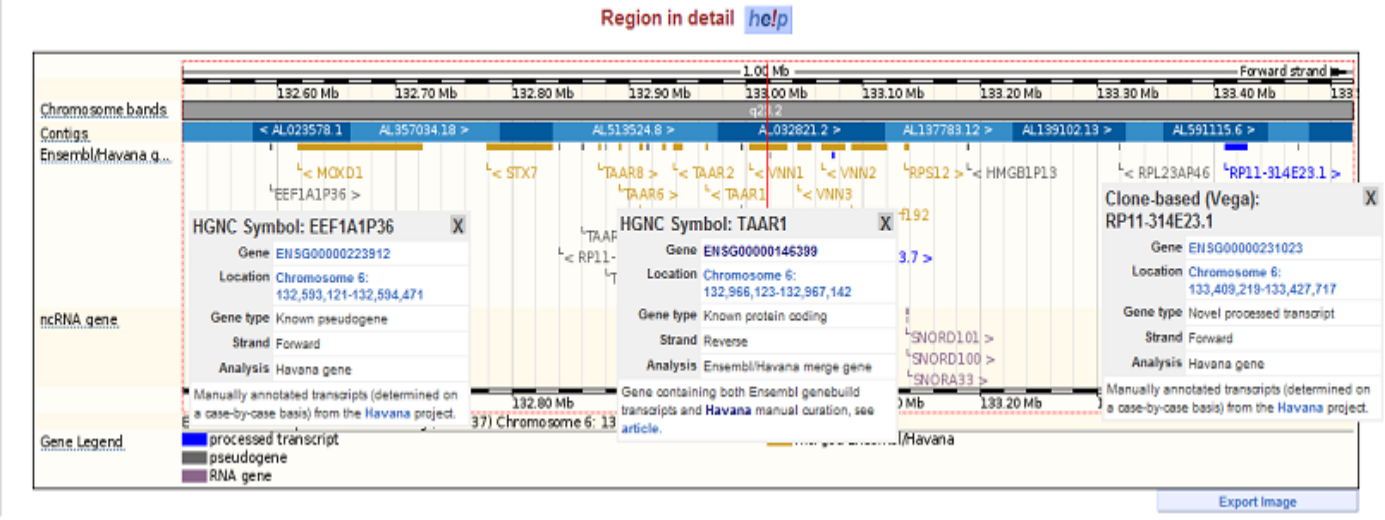


Рис.3 Информация о выбранных элементах

Далее, как видно из рис.4 , можно получить более детальную информацию:

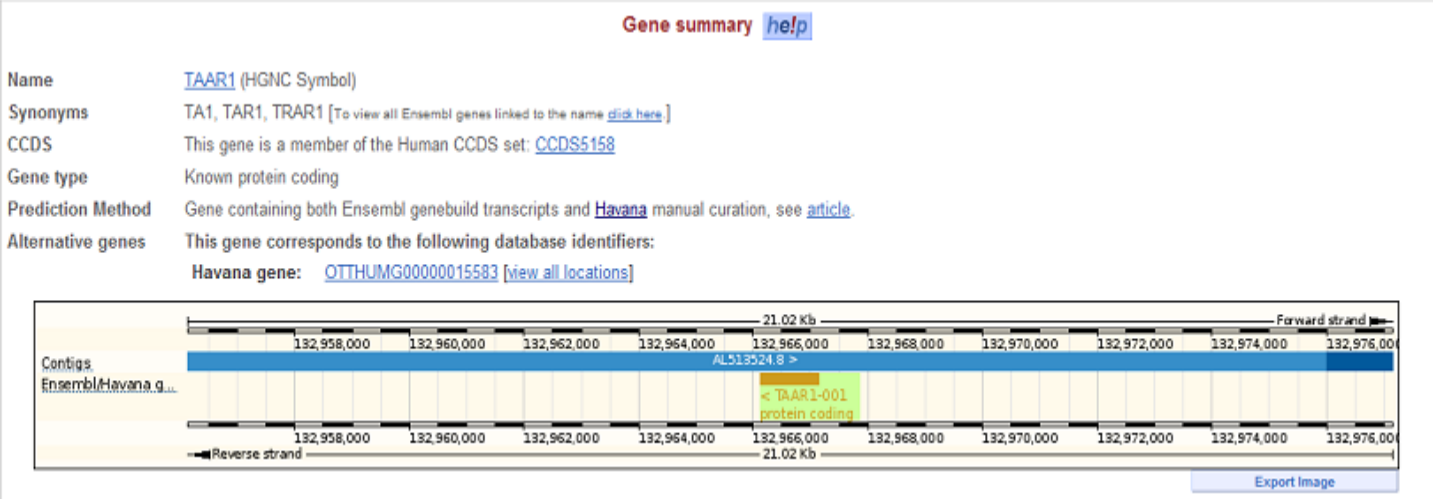


Рис.4 Детальная информация о выбранном элементе

С помощью Ensembl можно получить информацию о локусах [4], генах и повторах [5], попадающих в заданную область. К сожалению, информация о генах недостаточно подробна – нет сведений о кодирующих (экзонах) и некодирующих (интронах) участках гена, а эти сведения могут быть довольно полезны.

UCSC Table Browser

UCSC (University of California, Santa Cruz) Table Browser [6] обеспечивает доступ к огромному количеству данных о геномах различных организмов. С помощью данного сервиса можно указать геном, хромосому и интересующий нас диапазон, а также тип информации, которую мы хотим получить. Недостатками данного сервиса является большое число настроек, которые нужно задать, ограничения в типе получаемой информации (нельзя задать несколько интересующих нас таблиц, по которым будет вестись поиск) и формат выдаваемого результата.

На рис.5 показан пример использования Table Browser для получения информации о генах, пересекающихся с участком 22-ой хромосомы человеческого генома, начиная с 33265750 и заканчивая 33266750 позицией:

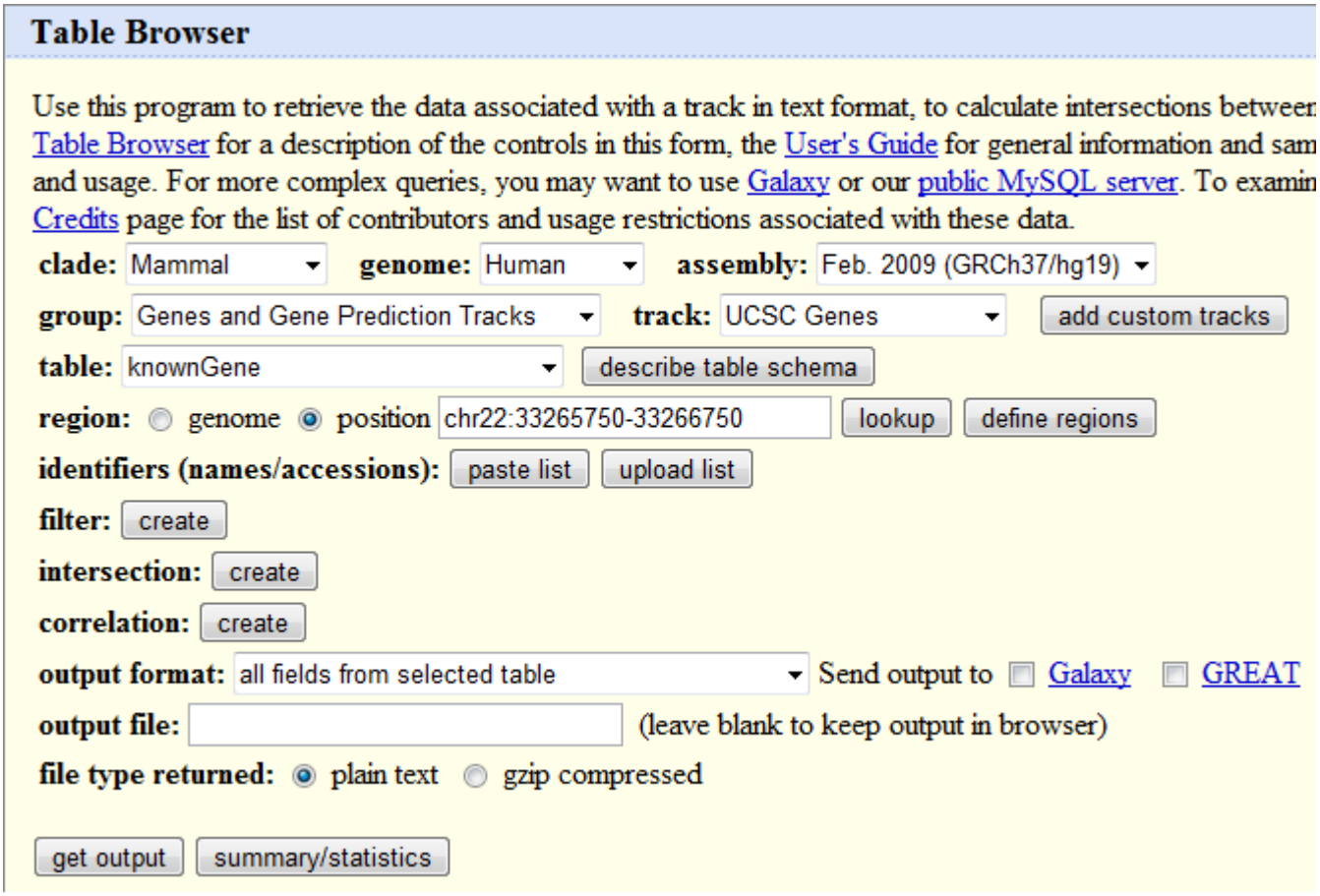


Рис.5 Пример использования Table Browser

Как видно из рис.6 , получаемые результаты иногда трудно разбирать:

#name	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	exonCount	exonStarts	exonEnds	
uc003amx.2	chr22	-	32908541	33402809	32909678	33402647	13	32908541		
uc003amy.2	chr22	-	32908541	33402809	32909794	33402647	12	32908541		
uc003amz.2	chr22	-	32908541	33454377	32909678	33402647	14	32908541		

Рис.6 Результат работы Table Browser

С помощью Table Browser можно получить подробную информацию о генах вместе с их кодирующими и некодирующими участками, о локусах, повторах и SNP многих организмов, но для каждого вида элементов нужно создавать отдельный запрос, а также недостатком является то, что полученные результаты нужно в дальнейшем обработать, чтобы они были представлены в удобном виде.

UCSC Genome Browser

UCSC (University of California, Santa Cruz) Genome Browser [7] позволяет получать информацию о заданном участке хромосомы, предоставляя пользователю визуализацию заданного участка. Картинку можно увеличивать и уменьшать, тем самым расширяя или сужая диапазон, тогда участок будет показан с менее или более подробной информацией о содержащихся в нем элементах соответственно. Для каждого элемента на картинке можно получить его аннотацию, в частности, доступны аннотации генов и повторов. Но для получения информации о кодирующих и некодирующих участках генов придется воспользоваться Table Browser.

На рис.7 представлен пример отображения информации об участке с позиции 33265750 по 33266750 21-ой хромосомы человеческого генома:

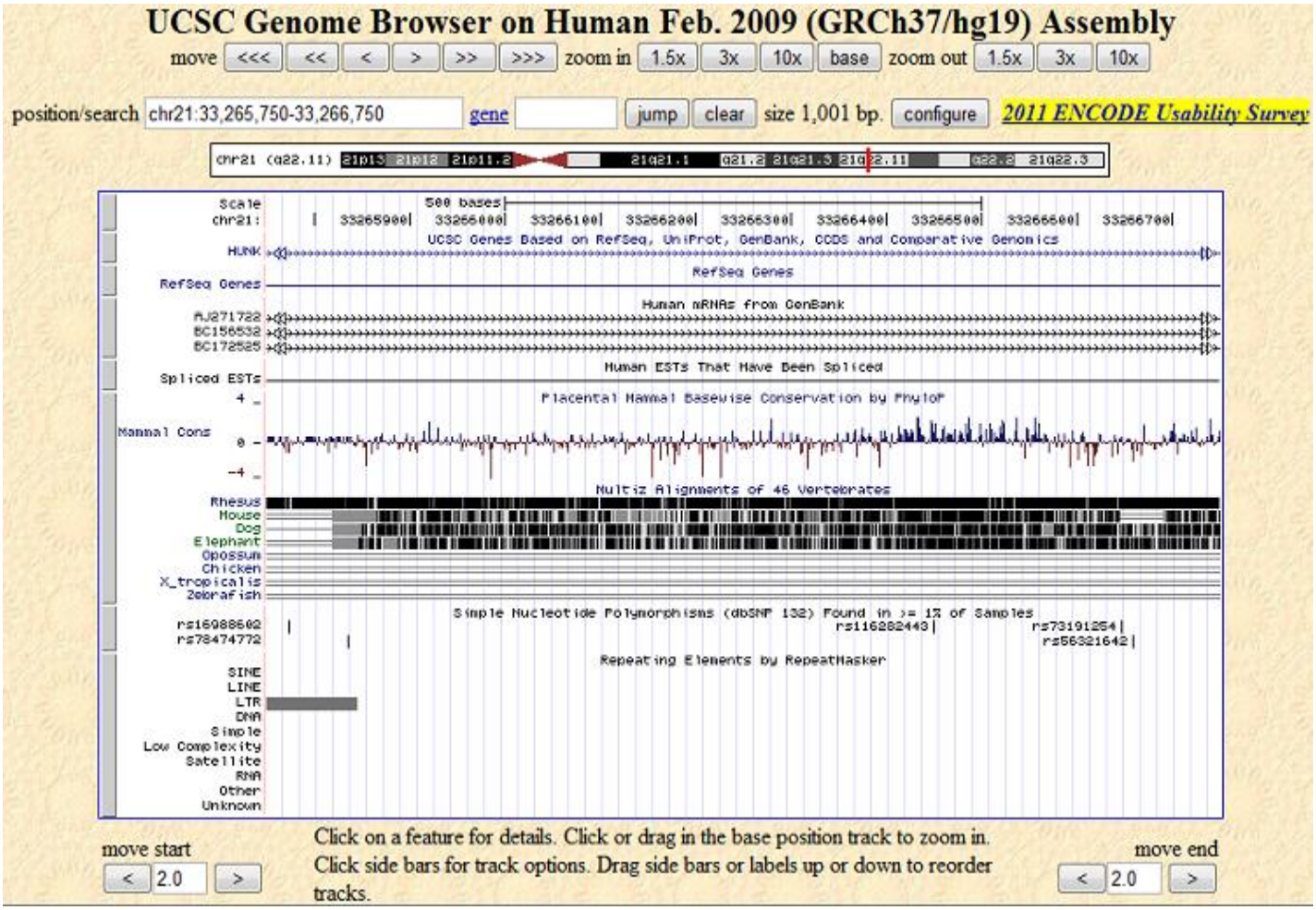
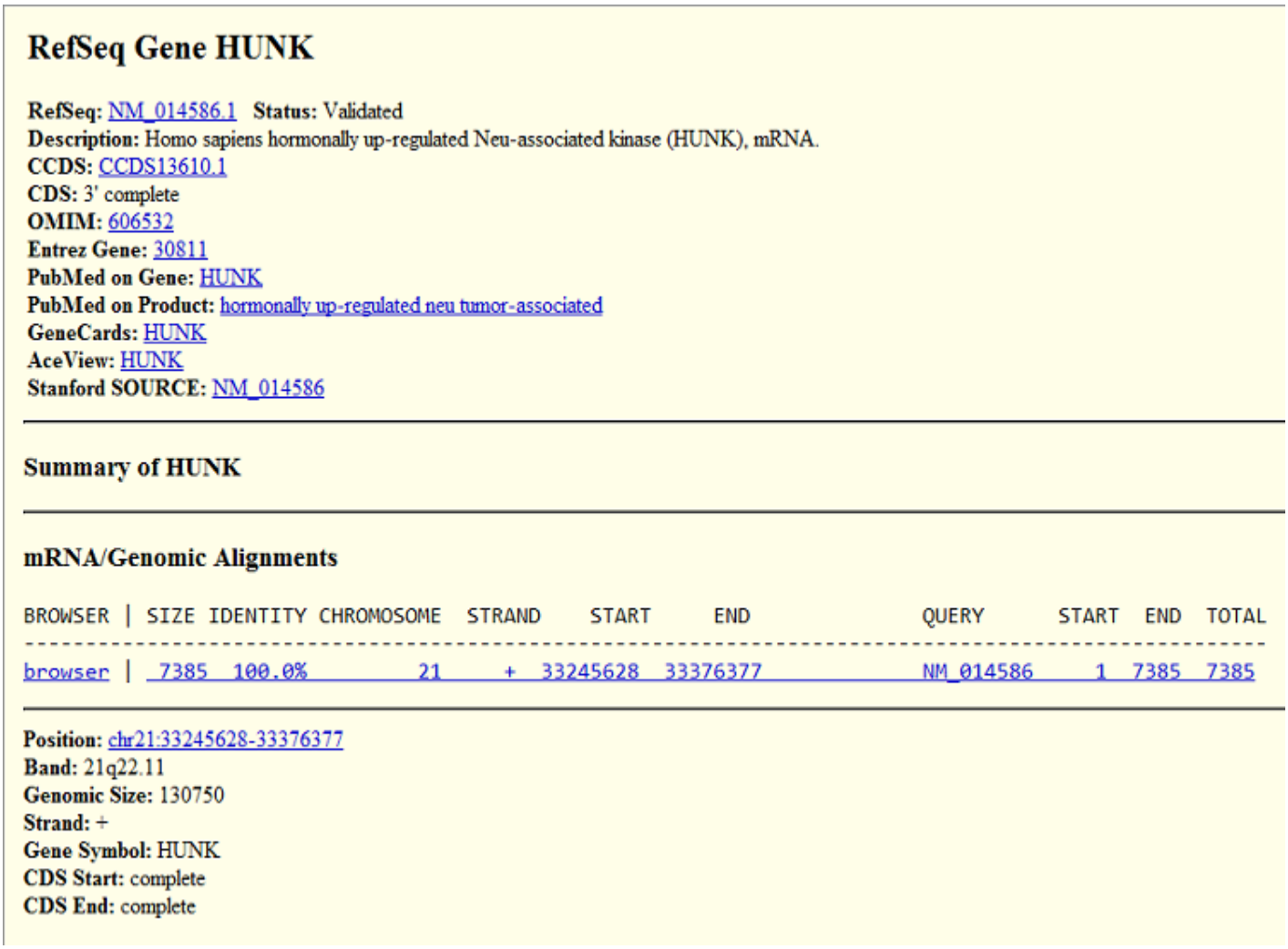


Рис.7 Результат работы Genome Browser

На рис.8 показан пример аннотации гена HUNK:



RefSeq Gene HUNK

RefSeq: NM_014586.1 **Status:** Validated
Description: Homo sapiens hormonally up-regulated Neu-associated kinase (HUNK), mRNA.
CCDS: CCDS13610.1
CDS: 3' complete
OMIM: 606532
Entrez Gene: 30811
PubMed on Gene: HUNK
PubMed on Product: hormonally up-regulated neu tumor-associated
GeneCards: HUNK
AceView: HUNK
Stanford SOURCE: NM_014586

Summary of HUNK

mRNA/Genomic Alignments

BROWSER	SIZE	IDENTITY	CHROMOSOME	STRAND	START	END	QUERY	START	END	TOTAL
browser	7385	100.0%	21	+	33245628	33376377	NM_014586	1	7385	7385

Position: chr21:33245628-33376377
Band: 21q22.11
Genomic Size: 130750
Strand: +
Gene Symbol: HUNK
CDS Start: complete
CDS End: complete

Рис.8 Аннотация гена HUNK

Итоги обзора существующих решений

В результате исследования работы вышеописанных сервисов было принято решение добавить в проект Genome Query возможность получения информации о последовательности нуклеотидов по ее местоположению в геноме человека, этому поспособствовало несколько причин.

Каждый из рассмотренных сервисов имеет свои преимущества, но также и недостатки. В Ensembl и Genome Browser отсутствует информация об экзонах и интронах, она присутствует только в таблицах, использующихся в Table Browser, но результаты работы данного сервиса трудны для понимания, иногда их необходимо обработать и представить в более наглядном виде. Ensembl и Genome Browser представляют результат работы в графическом виде, что очень удобно для пользователя, но если результаты будут использоваться для дальнейших исследований и если запросов много, то желательно иметь возможность автоматизации обработки полученной информации, а также иметь возможность быстро использовать полученные результаты для других запросов и исследований. Для этих целей более подходящими являются результаты в текстовом виде. Table Browser выдает результаты в текстовом виде, но процесс задания настроек запроса нетривиален.

В проекте «Genome Query» разработаны и реализованы алгоритмы для точного и неточного поиска нуклеотидных последовательностей, а также решается ряд других задач. Проект предоставляет веб-сервис и web-API, также ведётся разработка библиотеки реализованных алгоритмов для поиска последовательностей. Было решено добавить возможность получения информации о последовательности нуклеотидов по ее местоположению в геноме как, чтобы иметь более мощное средство для исследований, и пользователи могли бы использовать всего один инструмент для решения своих задач. Так как проект предоставляет web-API, возможно решение задач с большим количеством запросов, которые не нужно будет вводить вручную.

Данные о локусах, повторах, а также генах и их кодирующих и некодирующих участках были взяты из таблиц, использующихся в Table Browser, поскольку в них содержится огромное количество информации, взятое из GenBank.

Реализация собственного решения поставленной задачи

Данные о различных элементах считываются из таблиц, при этом создаются экземпляры соответствующих классов элементов (генов, повторов, локусов). Каждый такой объект имеет несколько атрибутов, в том числе начальную и конечную позиции на хромосоме, обозначающие местоположение данного элемента, номер хромосомы, имя элемента и спираль, на которой элемент находится, которые образуют класс Location:

```
public class Location {
   private int myStartIndex;
   private int myEndIndex;
   private Chromosome myChromosome;
   private Strand myStrand;
   …
}
```

Количество объектов насчитывает сотни тысяч, в частности, несколько десятков тысяч генов, каждый из которых может содержать десятки экзонов и интронов. При этом, нуклеотидные последовательности, соответствующие элементам, могут пересекаться. Для того чтобы по позиции в геноме определить, какие элементы содержат данную позицию, был разработан следующий алгоритм:

Вводится структура: весь геном делится на хромосомы, а те, в свою очередь, на части длиной в миллион нуклеотидов (точнее, в 2^20), при этом последние части имеют другую длину, зависящую от количества нуклеотидов в хромосоме. Тем самым мы получаем массив из 25 элементов (по числу хромосом), каждый элемент которого представляет собой разбиение хромосомы на части, а каждая часть отвечает за определенный участок генома. Эти участки будут заполняться информацией об элементах (экзонах, интронах, локусах, повторах), пересекающих данный участок. На рис. 9 показано начальное состояние структуры на примере одной хромосомы:

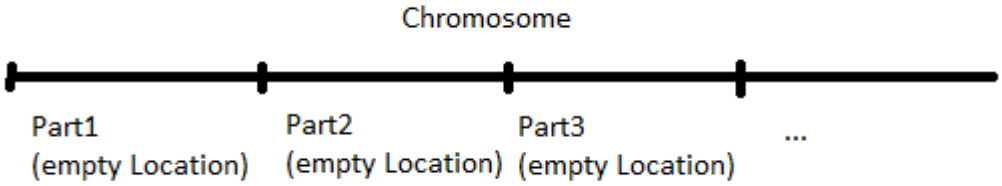


Рис.9 Начальное состояние структуры

При считывании геномной информации из таблиц, каждый созданный объект учитывается в структуре и разбивает ее на подчасти, что продемонстрировано на рис.10:

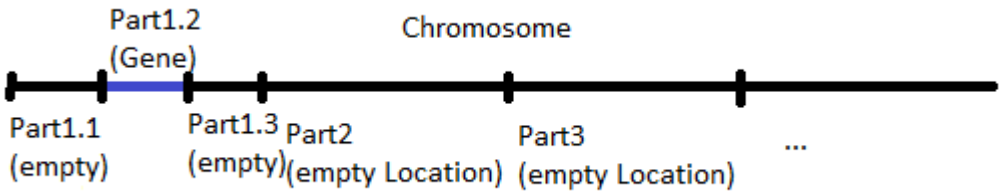


Рис.10 Добавление информации об элементе

При дальнейшем считывании информации из таблиц происходит добавление информации о них в структуру, части разделяются на непересекающиеся геномные участки, содержащие список элементов, пересекающих этот участок. На рис.11 показан пример такого дробления (считывается информация о Gene1, при этом в структуре уже имелся участок, который содержал информацию только о Gene2; в результате получаем три непересекающихся участка):

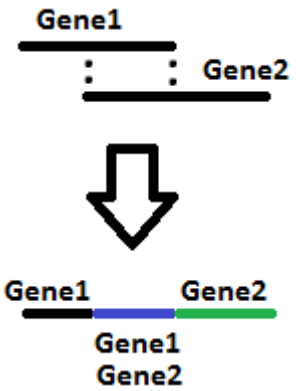


Рис.11 Дробление участка при считывании новой информации

В результате таких действий мы получаем структуру, состоящую из большого числа непересекающихся участков (можно сказать, диапазонов). При запросе информации для какой-либо позиции заданной хромосомы из структуры выбирается нужный элемент массива, соответствующий заданной хромосоме; далее, по значению хэш-функции от заданной позиции (в данном случае хэш-функция представляет собой целую часть от деления числа, соответствующего позиции, на 2^20) выбирается нужный участок, а в нём, с помощью бинарного поиска, выбирается нужный участок, в котором содержится список элементов, его пересекающих. Тем самым мы получаем интересующую нас информацию.

В настоящее время результат работы выдается в виде списка имен элементов, пересекающих заданный диапазон или позицию. Поскольку результат в виде текста, а не картинки, его можно использовать для последующих запросов и исследований, при этом работу можно будет автоматизировать. Также с помощью Genome Query можно получить более детальную информацию об элементах или, к примеру, посмотреть некоторую нуклеотидную последовательность. Примеры работы сервиса продемонстрированы на рис.12-14:

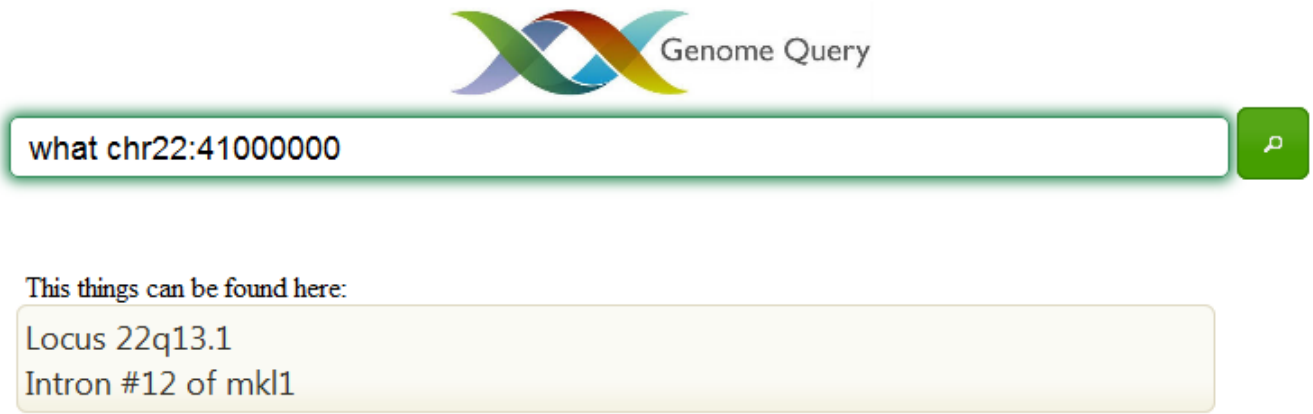


Рис12. Пример работы сервиса

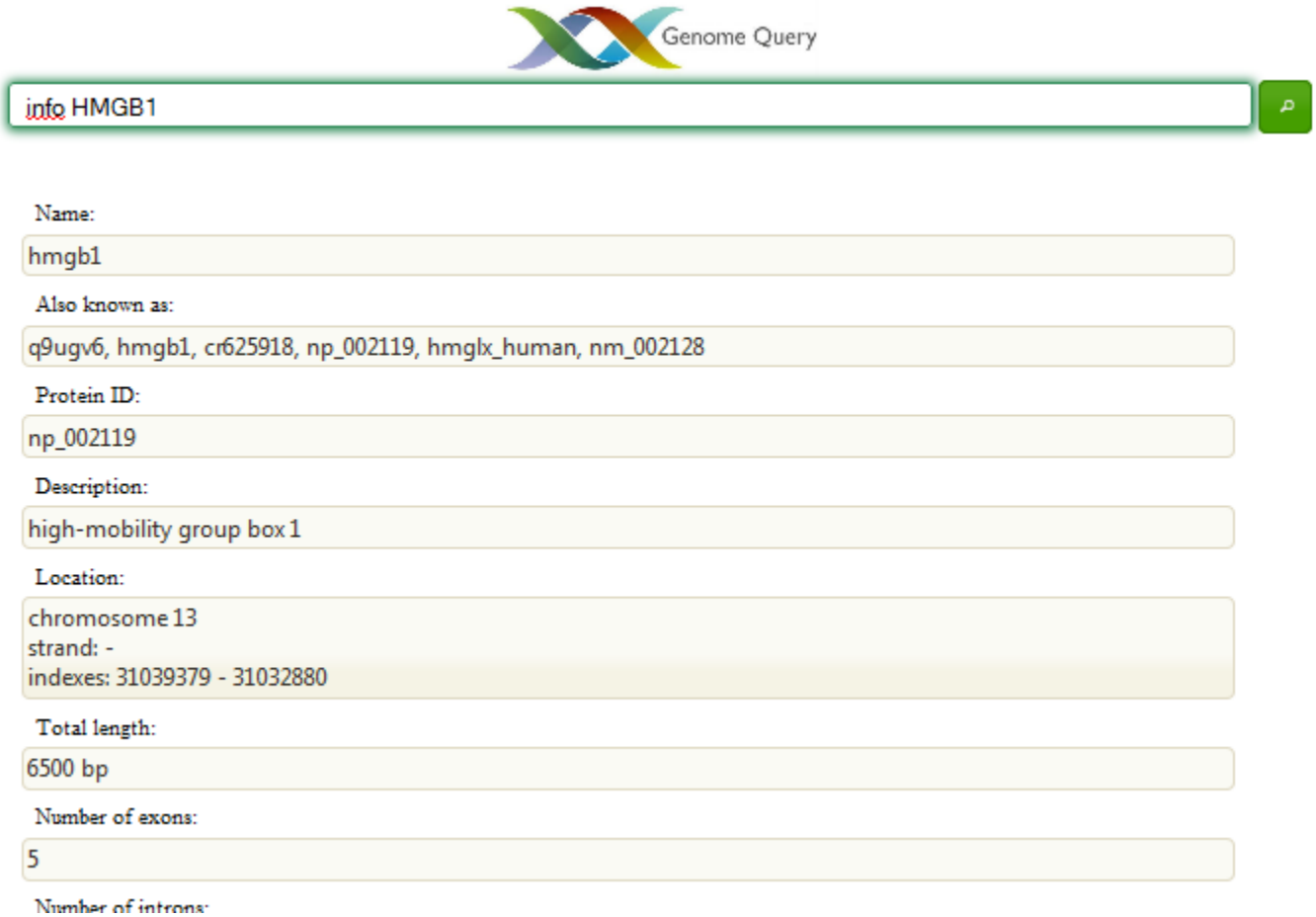


Рис.13 Детальная информация о гене

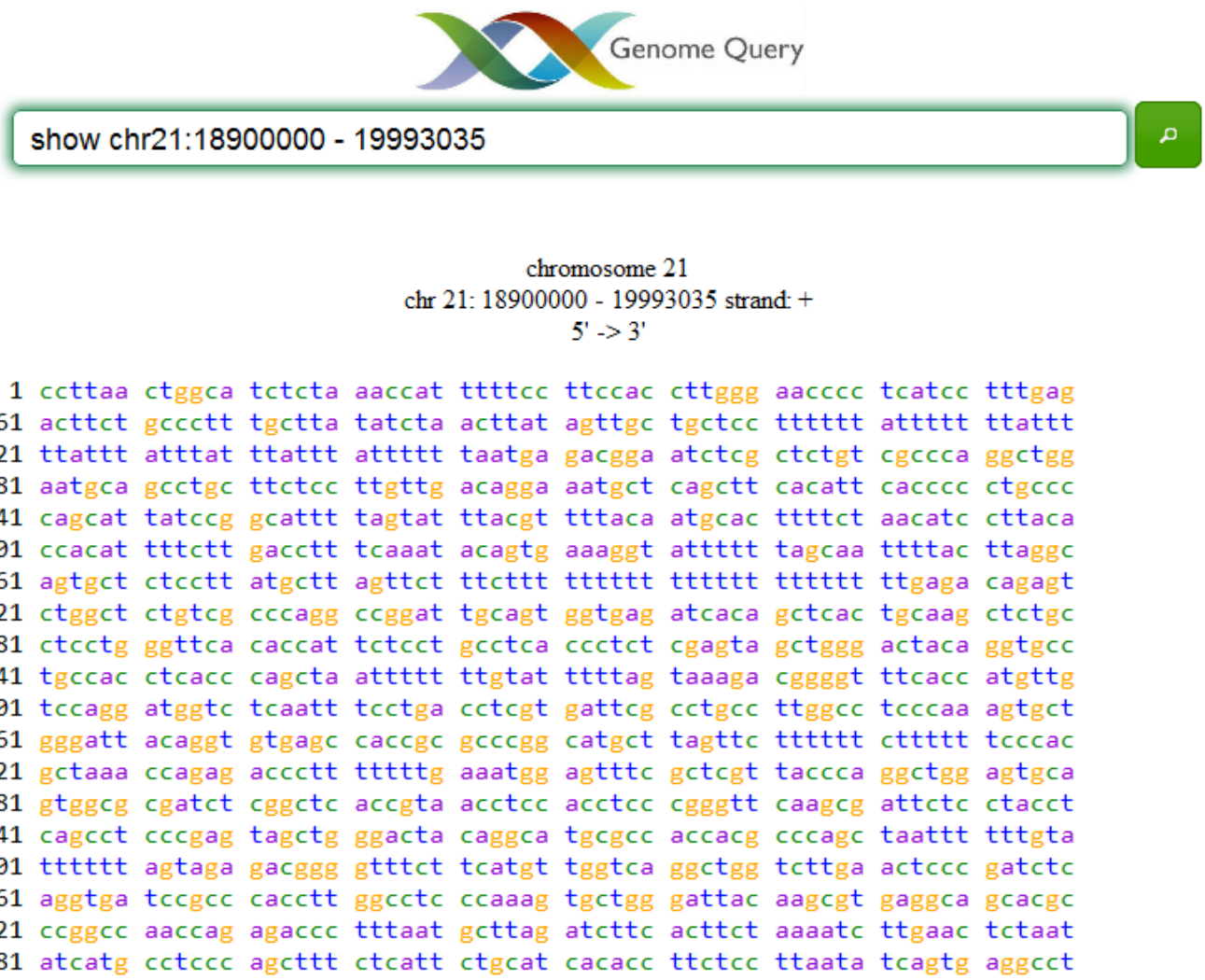


Рис.14 Отображение последовательности нуклеотидов

Скорость работы алгоритма не была протестирована при большой нагрузке на сервер. На данном этапе можно лишь сказать, что считывание и создание структуры занимают несколько секунд, а поиск информации для диапазона длиной в целую хромосому – около 50 миллисекунд, в зависимости от хромосомы. В ближайшее время скорость алгоритма будет более тщательно протестирована, как и скорость работы всего сервиса.

Заключение

В данной работе был проведен обзор существующих решений задачи определения функционального смысла нуклеотидной последовательности по ее местоположению в геноме, разработан и реализован собственный алгоритм для решения поставленной задачи. Данная функциональность была добавлена в проект Genome Query.

В дальнейшем планируется искать возможности для оптимизации алгоритма по памяти и времени, а также увеличить количество предоставляемой информации об участках генома. Также планируется реализация графического представления результатов работы для более удобного пользования сервисом.

Используемые источники

1. EMBL Nucleotide Databank
   http://www.ebi.ac.uk/embl/
2. Ensembl
   http://www.ensembl.org/index.html
3. Human Genome Project
   http://en.wikipedia.org/wiki/Human_Genome_Project
4. Локусы
   http://en.wikipedia.org/wiki/Locus_(genetics)
5. Повторы
   http://en.wikipedia.org/wiki/Human_genome#Repeat_elements
6. UCSC Table Browser
   http://genome.ucsc.edu/cgi-bin/hgTables?org=human
7. UCSC Genome Browser
   http://genome.ucsc.edu/cgi-bin/hgTracks?org=human