Для хранения больших объемов генетических данных необходимо использование баз данных или других специализированных хранилищ. Наиболее доступной базой данных является MySQL. Ее функциональные возможности (хранимые типы данных и возможные операции над ними) описаны производителем. В то время как производительность (объем занимаемого дискового пространства, скорость формирования базы данных и скорость выполнения запросов) необходимо тестировать на конкретном наборе данных.

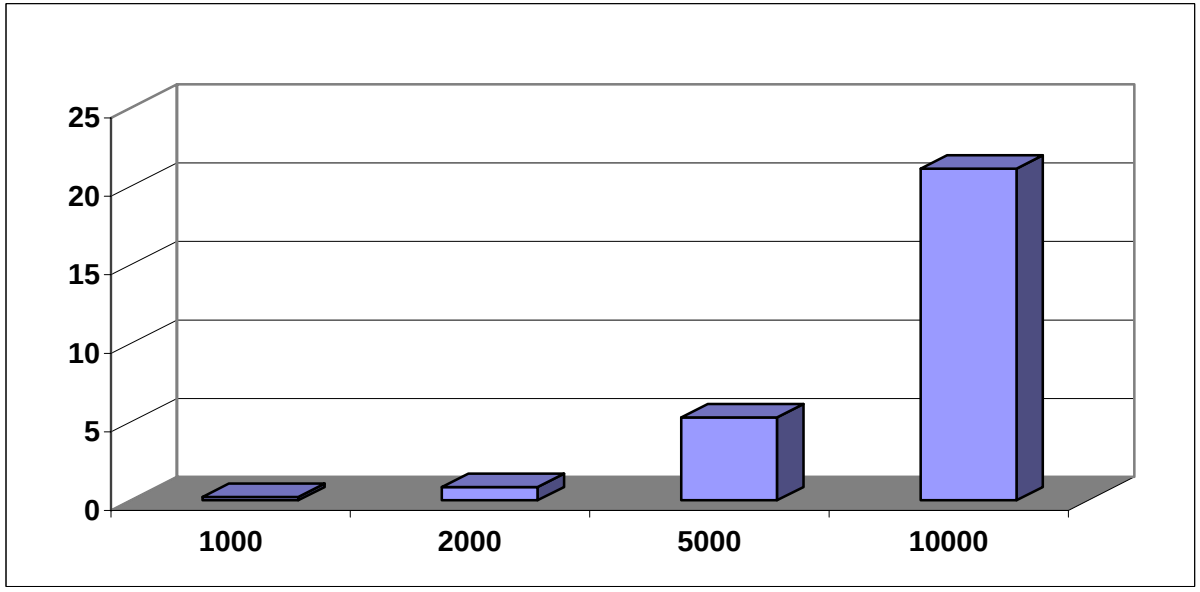
**Задача.** Определить производительность базы данных MySQL при хранении больших массивов генетических данных и сравнить ее с производительностью бинарного хранилища.

**Объект.** Генетические данные представлены в виде таблиц фенотипов, генотипов и индексов идентичности по происхождению (IBD). Размер этих таблиц зависит от числа людей ($N$), признаков ($P$) и локусов ($L$) и определяется для таблицы фенотипов как $P \times N$, генотипов – как $L \times N$, IBD – как $L \times N(N\text{-}1)/2$.

**Анализ производительности базы данных MySQL.**

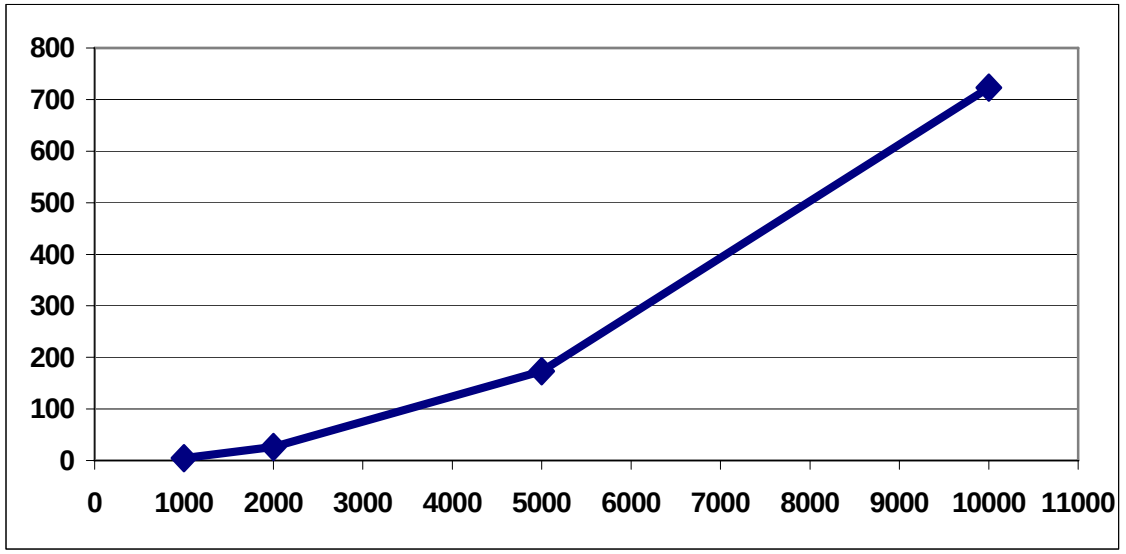
Объем дискового пространства (Гб), занимаемого базой данных MySQL при хранении матриц различного объема:

- объем выборки от 1.000 до 10.000 человек,
- число локусов 10.240,
- число признаков 100.



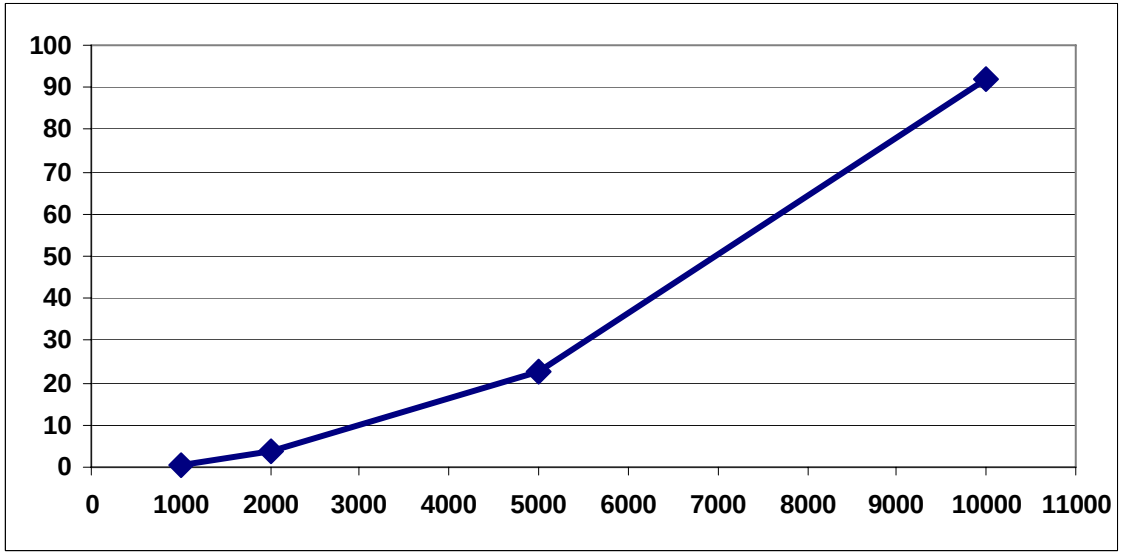
Время (мин.) создания базы данных MySQL при хранении матриц различного объема:

- объем выборки от 1.000 до 10.000 человек,
- число локусов 10.240,
- число признаков 100.




Время (мин.) извлечения всей информации из базы данных MySQL при хранении матриц различного объема:

- объем выборки от 1.000 до 10.000 человек,
- число локусов 10.240,
- число признаков 100.

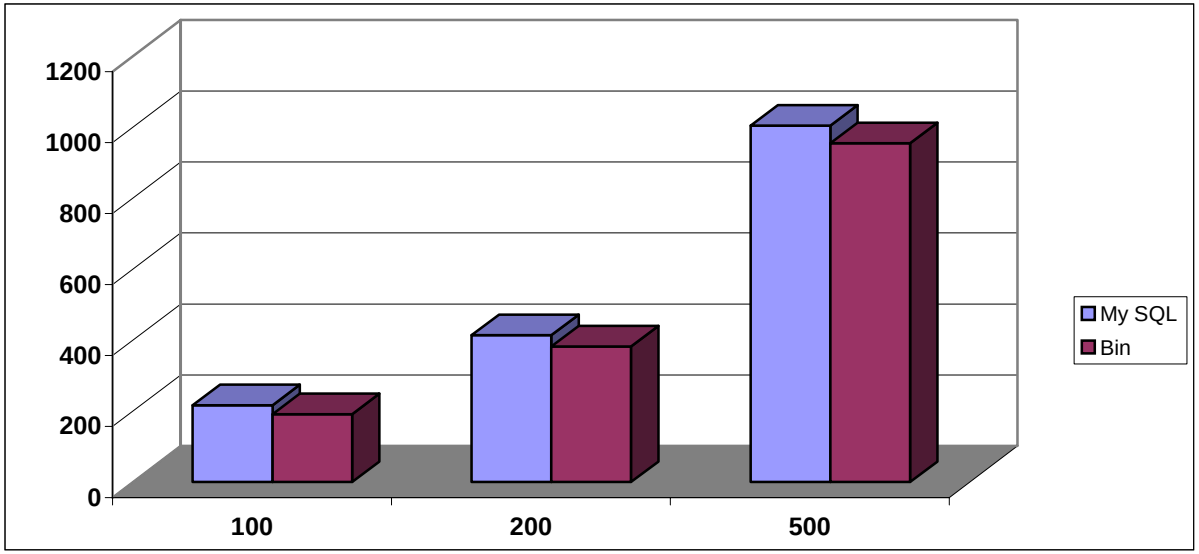


Тесты показали, что создание баз данных для хранения большого объема генетической информации требует больших ресурсов и растет нелинейно с ростом числа людей. Так как вся информация представлена однотипными данными и у каждого индивида имеется индивидуальный шифр, мы создали специализированное бинарное хранилище.

**Сравнение производительности MySQL и бинарного хранилища**

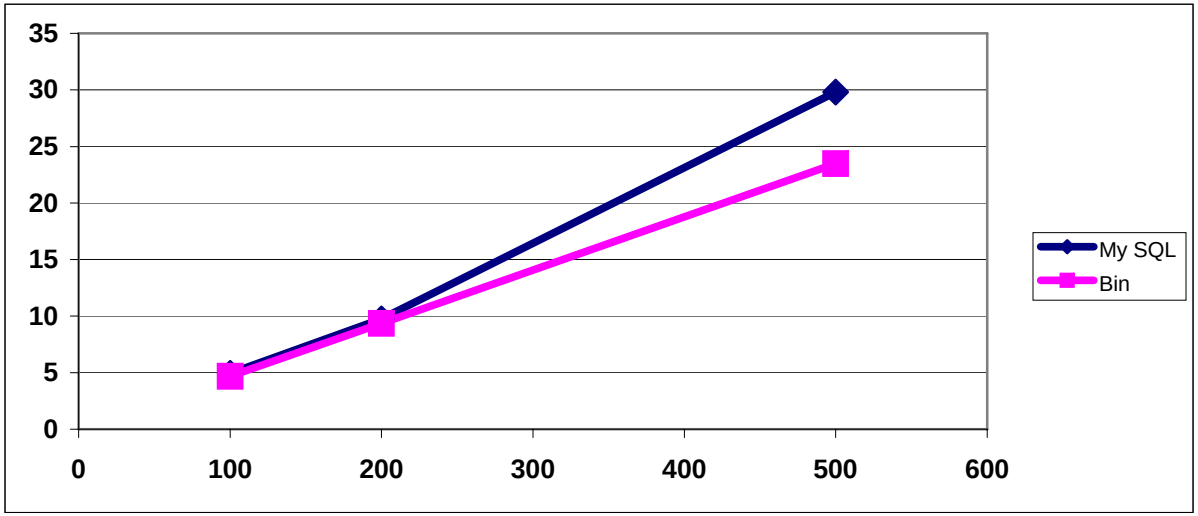
Объем дискового пространства (Гб), занимаемого базой данных MySQL и бинарного хранилища при хранении матриц различного объема:

- число локусов от 100 до 500
- объем выборки 1.000 человек
- число признаков 10.240.



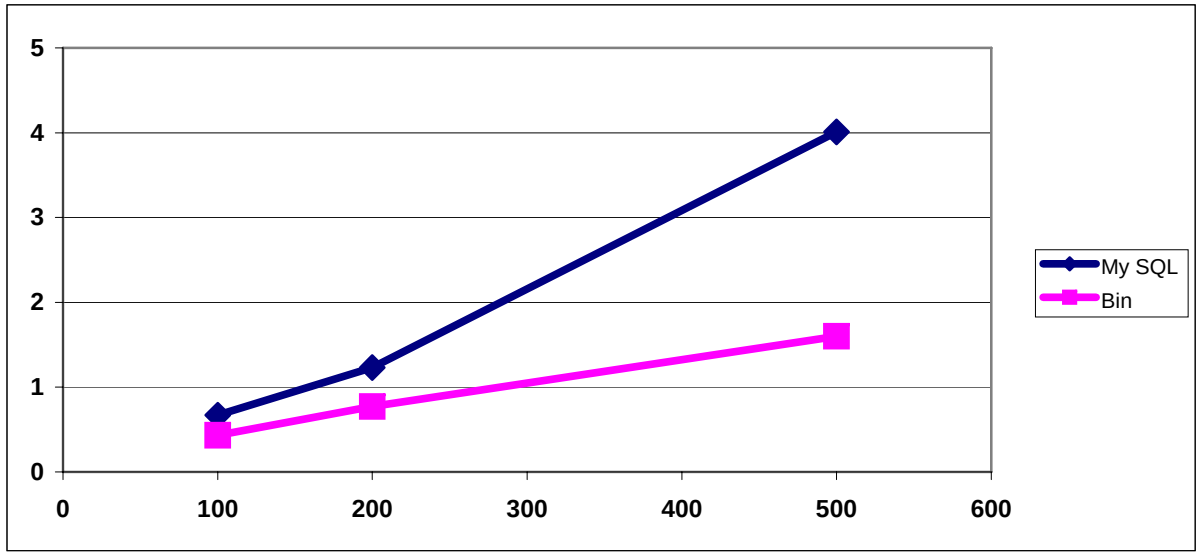

Время (мин.) создания базы данных MySQL и бинарного хранилища при хранении матриц различного объема:

- число локусов от 100 до 500
- объем выборки 1.000 человек
- число признаков 10.240.



Время (мин.) извлечения всей информации из базы данных MySQL и бинарного хранилища при хранении матриц различного объема:

- число локусов от 100 до 500
- объем выборки 1.000 человек
- число признаков 10.240.




**Вывод.** Сравнение MySQL и бинарного хранилища показало, что бинарное хранилище имеет ряд преимуществ:

- оно занимает меньшее дисковое пространство;
- благодаря меньшему объему, а также специфике данных, производительность такого хранилища выше.

Для выполнения запросов необходимо создать пакет программ. Для создания такого пакета лучше использовать языки программирования FORTRAN и C/C++, которые могут оперировать бинарными данными. Интерфейс для работы с базой данных предполагается сделать через Интернет страницы (html) и реализовать это на языках программирования PHP, Perl и возможно Java script.