

Создание веб-бота для автоматизации формирования запросов к порталу IMGТ / V-quest

Участники: Золотарев Андрей Владимирович, Чеблоков Александр Александрович

Руководитель: Бакин Евгений Александрович

Цель проекта

Существует большое множество различных веб-сайтов, предоставляющих возможность работы с интегрированными инструментами (BLAST, pfam, uniprot, phytozome).

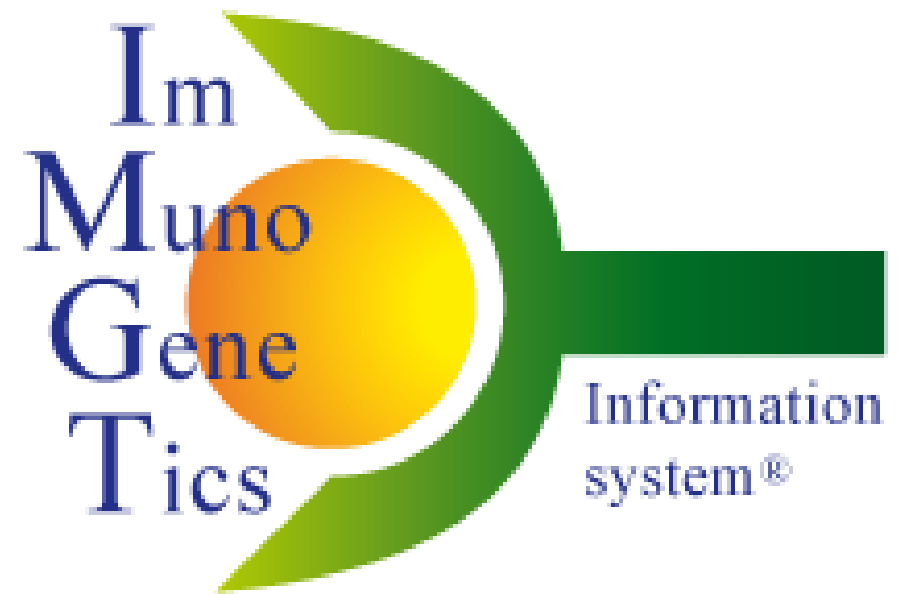
Проблемы:

- Ограничение по объему вводимых данных
- Рутинная настройка многочисленных опций

Целью проекта стала автоматизация запросов для одного из таких инструментов, а именно **IMGIT / V-quest**.

IMGT

IMGT[®], (the international ImMunoGeneTics information system[®]) - это высококачественный источник знаний по иммуногенетике и иммуноинформатике



<http://www.imgt.org>

IMGT / V-quest

IMGT / V-QUEST это интегрированный инструмент выравнивания для нуклеотидных последовательностей **иммуноглобулинов** и **рецепторов Т-клеток**.

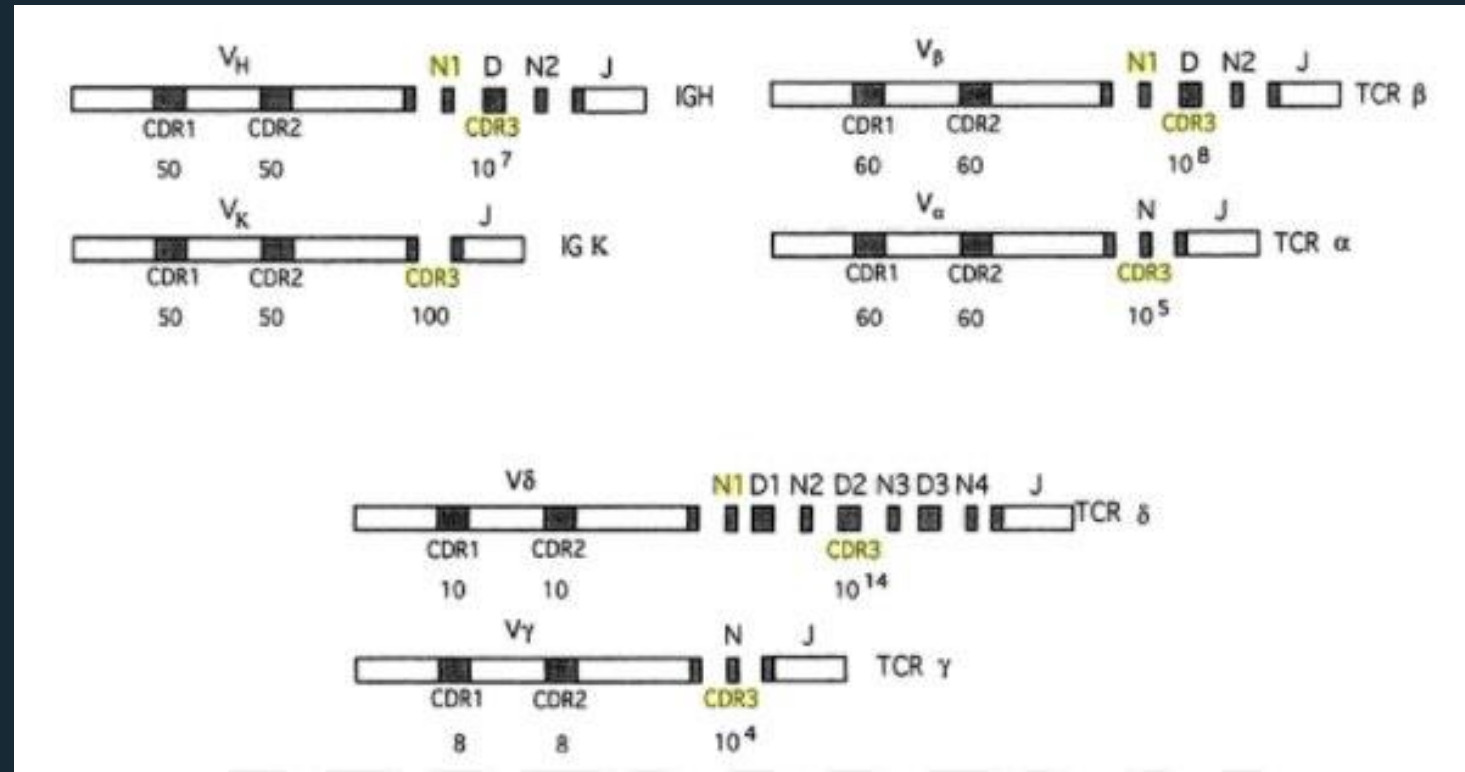


Рис. 1. Гены иммуноглобулинов с отмеченными зонами

Для формирования запроса необходимо:

1. Выбрать вид
2. Выбрать тип рецептора или локус
3. Ввести последовательности (не более 50 штук) в формате FASTA
4. Выбрать формат отчета
5. Выбрать другие опции отчета

Analyse your IG (or antibody) or TR nucleotide sequences

The list of the IMGT/V-QUEST reference directory sets to which your sequences can be compared is available in [here](#).

Human sequence sets to test IMGT/V-QUEST are available [here](#)

The screenshot shows the IMGT/V-QUEST web interface. It is divided into several sections:

- Species selection:** A dropdown menu labeled "Species" with a red box and the number "1" next to it.
- Receptor type or locus selection:** A dropdown menu labeled "Receptor type or locus" with a red box and the number "2" next to it.
- Sequence submission:** A section with a red box and the number "3" next to it. It contains two radio buttons: "Type (or copy/paste) your nucleotide sequence(s) in FASTA format" (selected) and "Or give the path access to a local file containing your sequence(s) in FASTA format". Below the second option are buttons for "Выберите файл" and "Файл не выбран", and "Start" and "Clear the form" buttons.
- Display results:** A section with a red box and the number "4" next to it. It features radio buttons for "A. Detailed view" (selected), "HTML" (selected), and "Text". It also includes dropdown menus for "Nb of nucleotides per line in alignments:" (set to 60) and "Nb of aligned reference sequences:" (set to 5).
- Results options:** A section with a red box and the number "5" next to it. It contains 14 numbered checkboxes for various analysis options, such as "Alignment for V-GENE", "V-REGION alignment", "Sequences of V-, V-J- or V-D-J- REGION", etc. At the bottom are links for "Check all", "Uncheck all", and "Default".

Рис. 2. Окно ввода данных для инструмента IMGT / V-QUEST

Вывод IMGT / V-quest

Label	Location/Qualifiers	CDR3-IMGT		
13. Annotation by IMGT/Automat				
V-D-J-REGION	1..355 /CDR_length="[8.8.11]" /FR_length="[25.17.38.11]" /nucleotide sequence gaggtgcagctgttggagctctggggaggcgtgggtccagcctgggaggtccctgagactc tcctgtatagcctctggattcaccttcagtagctatcctatgacctgggtccgccaggct ccaggcaaggggctggagtggtggcaagtatatcatatgacggaagtataaataaag gtagactccatgaagggccgactcaccatctccagagacaattccaagaacacgctgtat ttggaaatgaacagcctgacagctgaggacacggctgtgtattactgtgcgaggacagct ttctttaacgcctatgacttctggggccagggaaacctgtcacctctcctcag /translation EVQLLESGGGVVQPGRSLRLSCLASGFTFSSYPMTWVRQAPGKLEWVASISYDGSYKYK VDSMKGRLTISRDNKNTLYLEMNSLTAEDTAVYYCARTAFNAYDFWQGLTIVTVSS	289..321	/AA_IMGT="AA 105 to 117, AA 111, 112 are missing" /nucleotide sequence gcgaggacagctttctttaacgcctatgacttc /translation ARTAFFNAYDF	
V-REGION	1..293 /allele="Homsap IGHV3-30*04 F or Homsap IGHV3-30-3*03 F" /gene="Homsap IGHV3-30 or Homsap IGHV3-30-3" /nucleotide sequence gaggtgcagctgttggagctctggggaggcgtgggtccagcctgggaggtccctgagactc tcctgtatagcctctggattcaccttcagtagctatcctatgacctgggtccgccaggct ccaggcaaggggctggagtggtggcaagtatatcatatgacggaagtataaataaag gtagactccatgaagggccgactcaccatctccagagacaattccaagaacacgctgtat ttggaaatgaacagcctgacagctgaggacacggctgtgtattactgtgcgaggacagct /translation EVQLLESGGGVVQPGRSLRLSCLASGFTFSSYPMTWVRQAPGKLEWVASISYDGSYKYK VDSMKGRLTISRDNKNTLYLEMNSLTAEDTAVYYCARTAFNAYDFWQGLTIVTVSS	286..324	/in_frame /nucleotide sequence tgtgaggacagctttctttaacgcctatgacttctgg /translation CARTAFFNAYDFW	
FR1-IMGT	1..75 /AA_IMGT="AA 1 to 26, AA 10 is missing" /nucleotide sequence gaggtgcagctgttggagctctggggaggcgtgggtccagcctgggaggtccctgagactc tcctgtatagcctctggattcaccttcagtagctatcctatgacctgggtccgccaggct ccaggcaaggggctggagtggtggcaagtatatcatatgacggaagtataaataaag gtagactccatgaagggccgactcaccatctccagagacaattccaagaacacgctgtat ttggaaatgaacagcctgacagctgaggacacggctgtgtattactgtgcgaggacagct /translation EVQLLESGGGVVQPGRSLRLSCLASGFTFSSYPMTWVRQAPGKLEWVASISYDGSYKYK VDSMKGRLTISRDNKNTLYLEMNSLTAEDTAVYYCA	3'V-REGION	286..293	/nucleotide sequence tgtgag /translation CA
1st-CYS	64..66 /nucleotide sequence tgt /translation C	(N-D)-J-REGION	294..355	/codon_start=2 /nucleotide sequence gacagctttctttaacgcctatgacttctggggccagggaaacctgtcacctctcctc ag /translation TAFFNAYDFWQGLTIVTVSS
CDR1-IMGT	76..99 /AA_IMGT="AA 27 to 38, AA 31, 32, 33, 34 are missing" /nucleotide sequence ggattcaccttcagtagctatcct /translation GFTFSSYP	(N-D)-REGION	294..311	/codon_start=2 /nucleotide sequence gacagctttctttaacgc /translation TAFFN
FR2-IMGT	100..150 /AA_IMGT="AA 39 to 55" /nucleotide sequence atgacctgggtccgccaggctccaggcaaggggctggagtggtggcaagt /translation MTWVRQAPGKLEWVAS	N1-REGION	294..294	/nucleotide sequence g
		D-J-REGION	295..355	/nucleotide sequence acagctttctttaacgcctatgacttctggggccagggaaacctgtcacctctcctca g /translation TAFFNAYDFWQGLTIVTVSS
		D-REGION	295..307	/allele="Homsap IGHDS-18*01" /gene="Homsap IGHDS-18" /nucleotide sequence acagctttcttta /translation TAFF
		N2-REGION	308..311	/codon_start=3 /nucleotide sequence acgc
		5'J-REGION	312..324	/codon_start=2

Рис. 3. Пример вывода IMGT / V-quest

Метод решения: Selenium WebDriver

Selenium WebDriver – это программная библиотека для управления браузерами.

Библиотеки WebDriver доступны на языках:

- Java
- .Net (C#)
- Python
- Ruby
- JavaScript

Драйверы реализованы для браузеров:

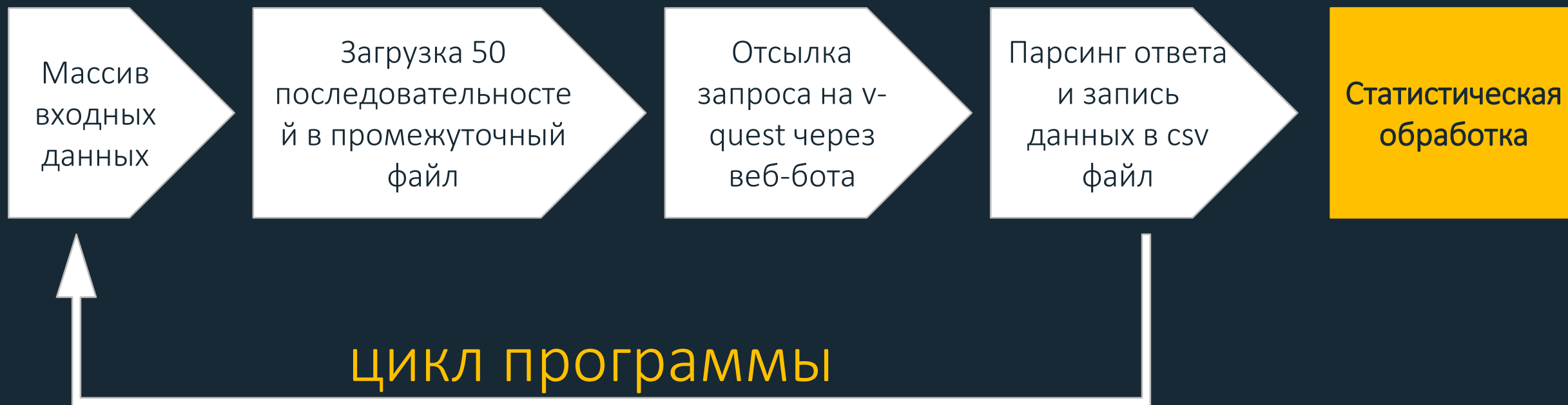
- Firefox
- InternetExplorer
- Safari
- Andriod
- iOS
- Chrome
- Opera



Проделанная работа:

1. Изучили основные команды API библиотеки Selenium WebDriver
2. Реализовали цикл, отбирающий нужные параметры, и загружающий на обработку по 50 последовательностей
3. Осуществили выгрузку данных с сервера, провели парсинг выходного файла и получили требуемые данные (длины CDR3 и N1 регионов) и сохранение в виде csv файла
4. Построили гистограммы распределения длин CDR3 и N1 регионов

Пайплайн программы



Парсинг

Парсинг ответов v-quest осуществлялся при помощи данного регулярного выражения:

'>(P<ID>[A-Z]{2}[d]{6}\.[d]{1}) #Поиск ID записи

[s\S]*? #Пропуск всех символов до следующего элемента

CDR3-IMGT[s]* #Поиск заголовка блока содержащего информацию о CDR3 регионе

(P<CD_start>[d]+)\.\.(P<CD_stop>[d]+) #Получение начала и конца CDR3

[s\S]*? #Пропуск всех символов до следующего элемента

N1-REGION[s]* #Поиск заголовка блока содержащего информацию о N1 регионе

(P<N1_start>[d]+)\.\.(P<N1_stop>[d]+)' #Получение начала и конца N1

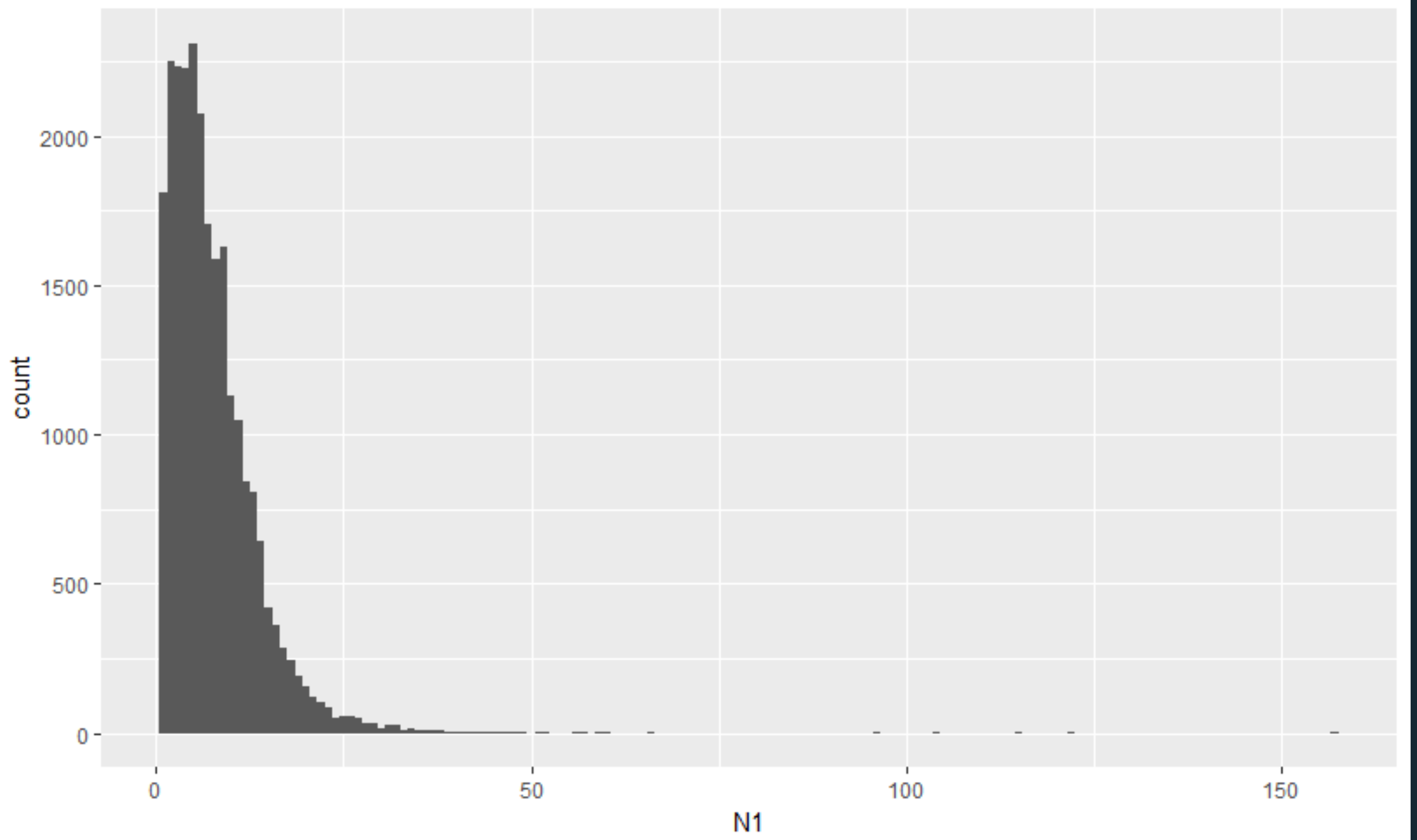


Рис. 4. Гистограмма распределения длины N1 региона

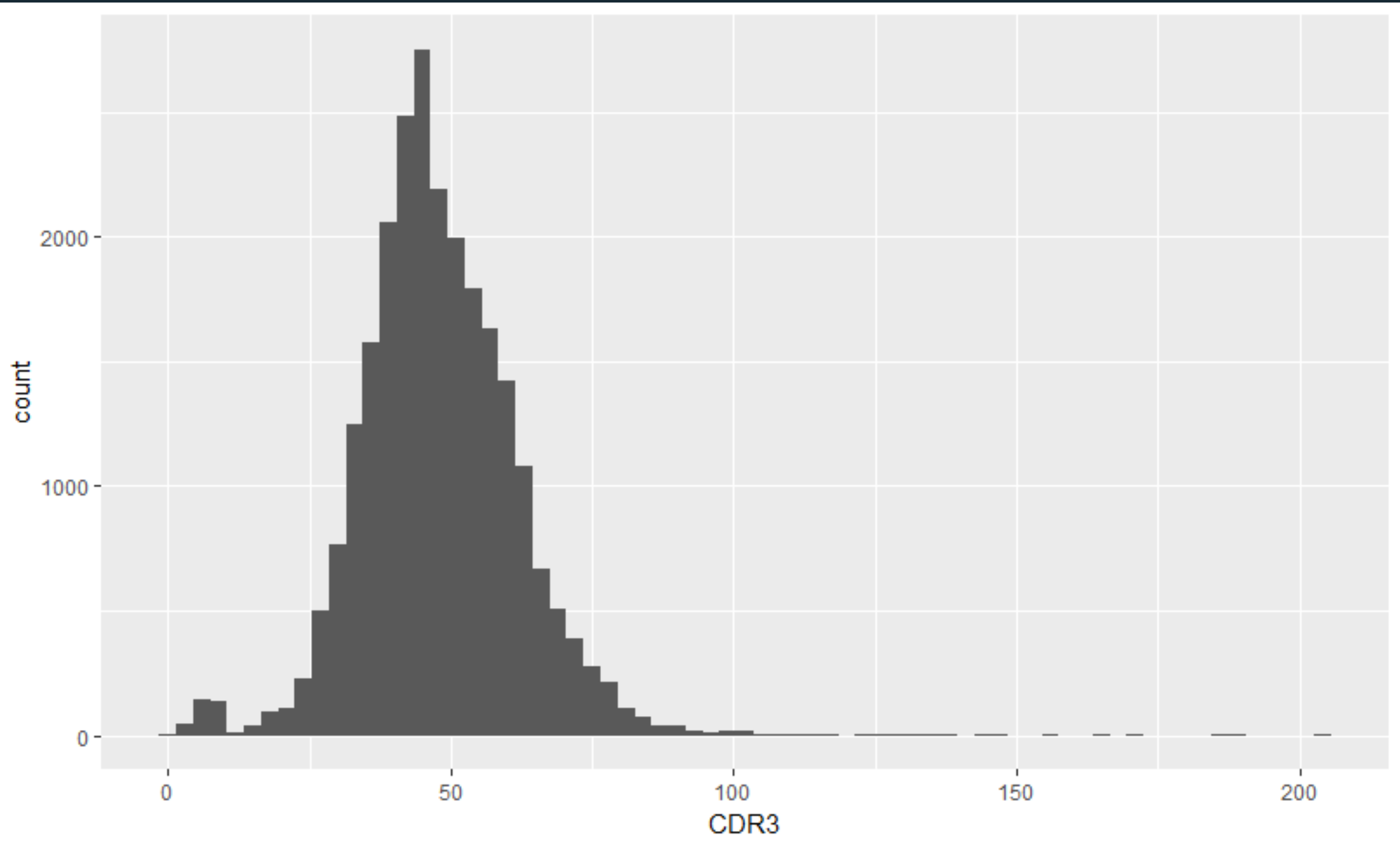


Рис. 5. Гистограмма распределения длины CDR3 региона

Заключение

1. Сделан веб-бот для автоматизированной отправки запросов на v-quest и последующего парсинга ответов
2. Обработан массив данных и построены гистограммы распределения длин регионов N1 и CDR3

Планы на будущее

- Реализация графического интерфейса
- Ускорение работы
- Расширение функционала в части выбора дополнительных опций для анализа и выгрузки полученных данных

Спасибо за внимание!