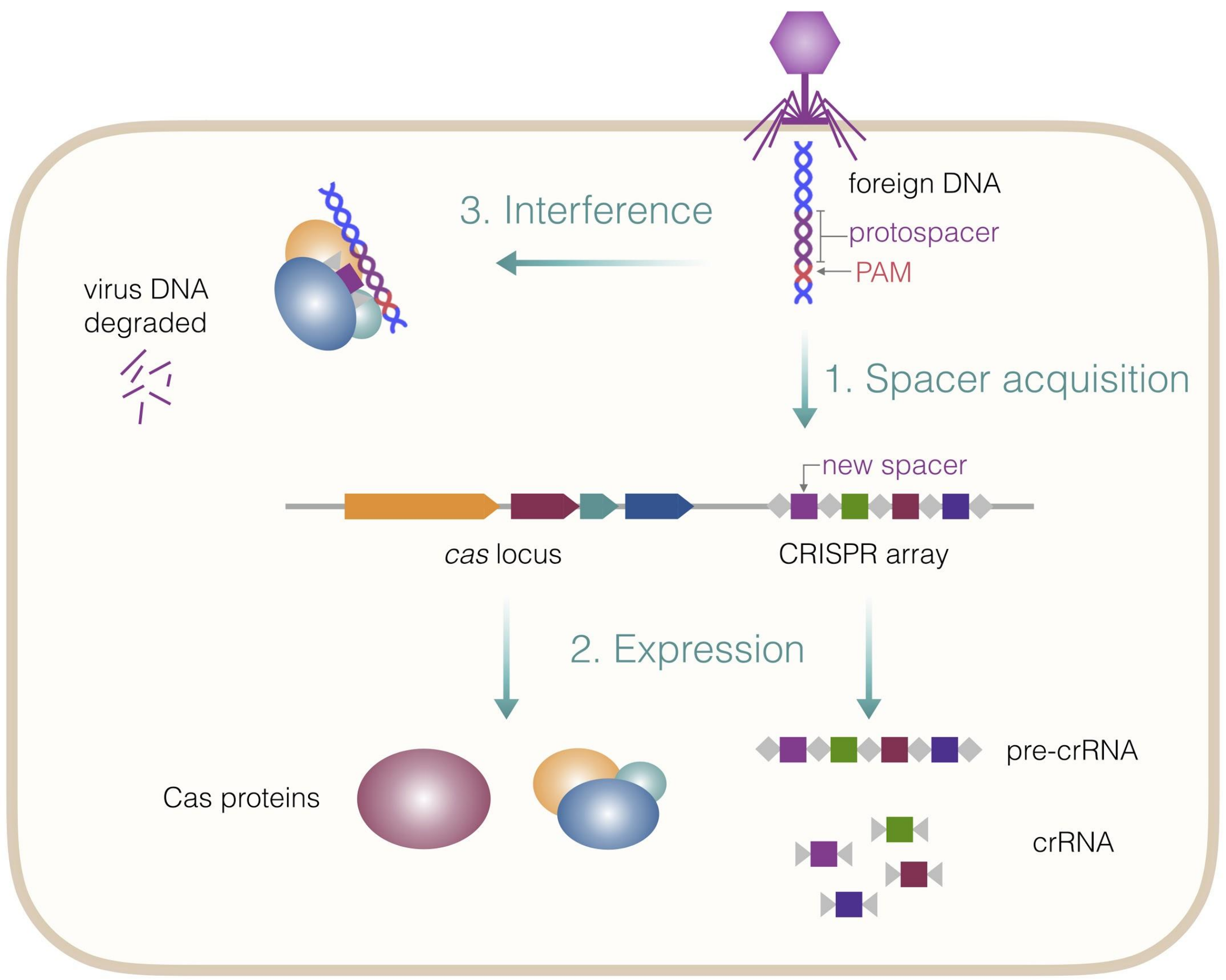
KU LEUVEN

CRISPR Exposed

A project by: Hamed Borhani, Mohamedhakim Elakhrass, Yi Ming Gan, Cedric Lood
Supervisors: Prof. Jan Aerts, Prof. Rob Lavigne

Introduction



Clustered, Regularly Interspaced, Short Palindromic Repeats (CRISPR) and CRISPR-associated proteins (Cas proteins) form a complex bacterial/archaeal immune response system that mitigates foreign DNA activity [1] . This immune-like system works by capturing short signatures of invading DNA and inserting them into the genome of the organism in regions known as CRISPR arrays. These arrays consist of the captured elements, known as spacers, which are separated by similarly sized, conserved DNA sequences known as repeats.

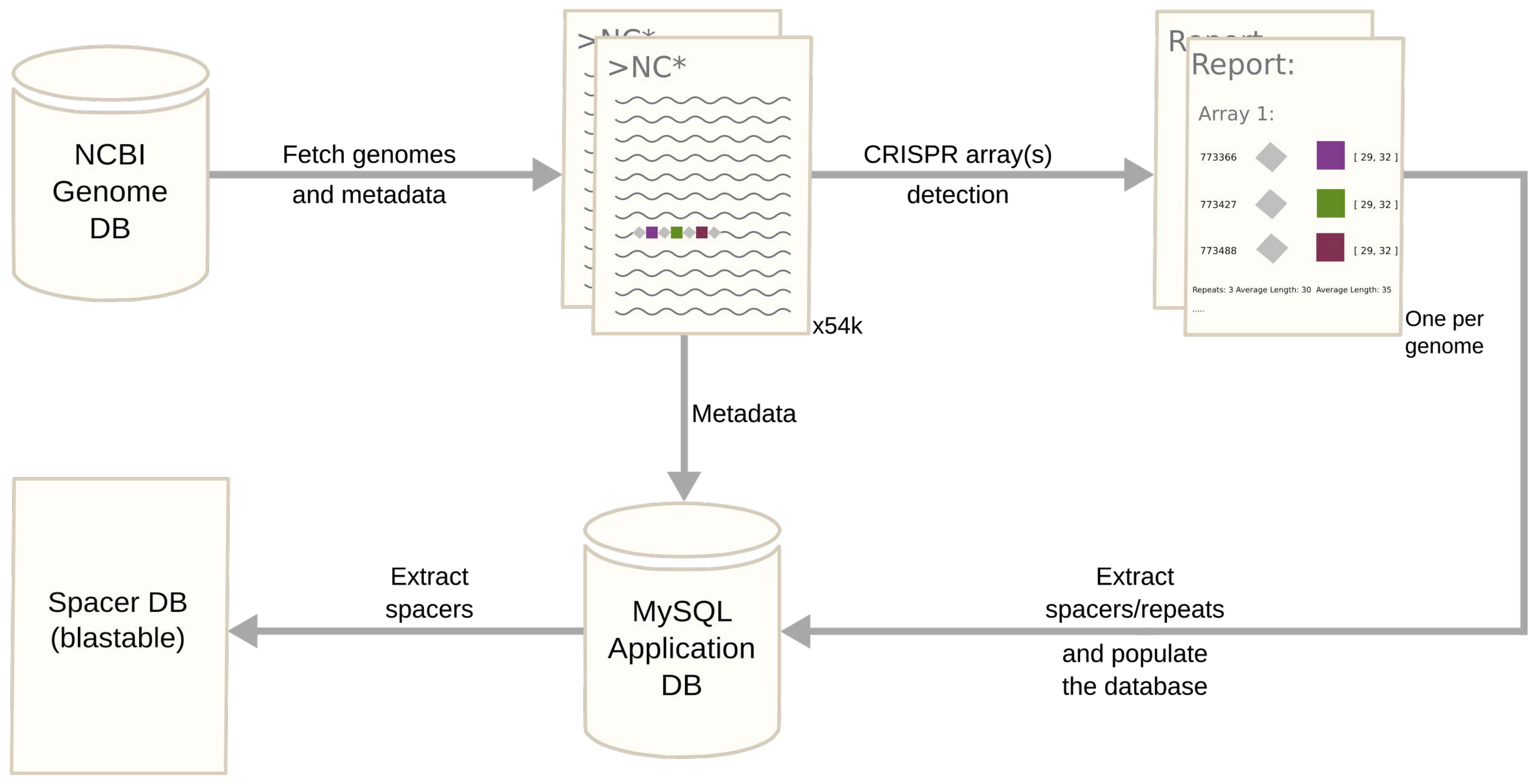
The arrays are expressed and processed into CRISPR RNA (crRNA). crRNA, together with Cas proteins, form an active complex that patrols the cell. If an invader DNA element with a similar signature is encountered by the complex, it will be degraded and its activity will be prevented.

The main objective of this project was to make a large database of CRISPR elements from all the bacterial/archaeal genomes available in the NCBI genome database. The database was made accessible via a web application. The associated website exposes the CRISPR details of the strains, along with a number of services, such as blasting against spacer elements, CRISPR array detection, and elements of data visualization.

Data processing

The first step of the data processing pipeline was to fetch all possible bacterial/archaeal genomes from NCBI genome database. The second step was to detect all CRISPR arrays (i.e spacers and repeats) in each genome. For this, the CRT software [2] was applied to the fetched genomes. A custom parser was implemented to extract the spacers and repeats from each CRT report and structure them into a JSON file.

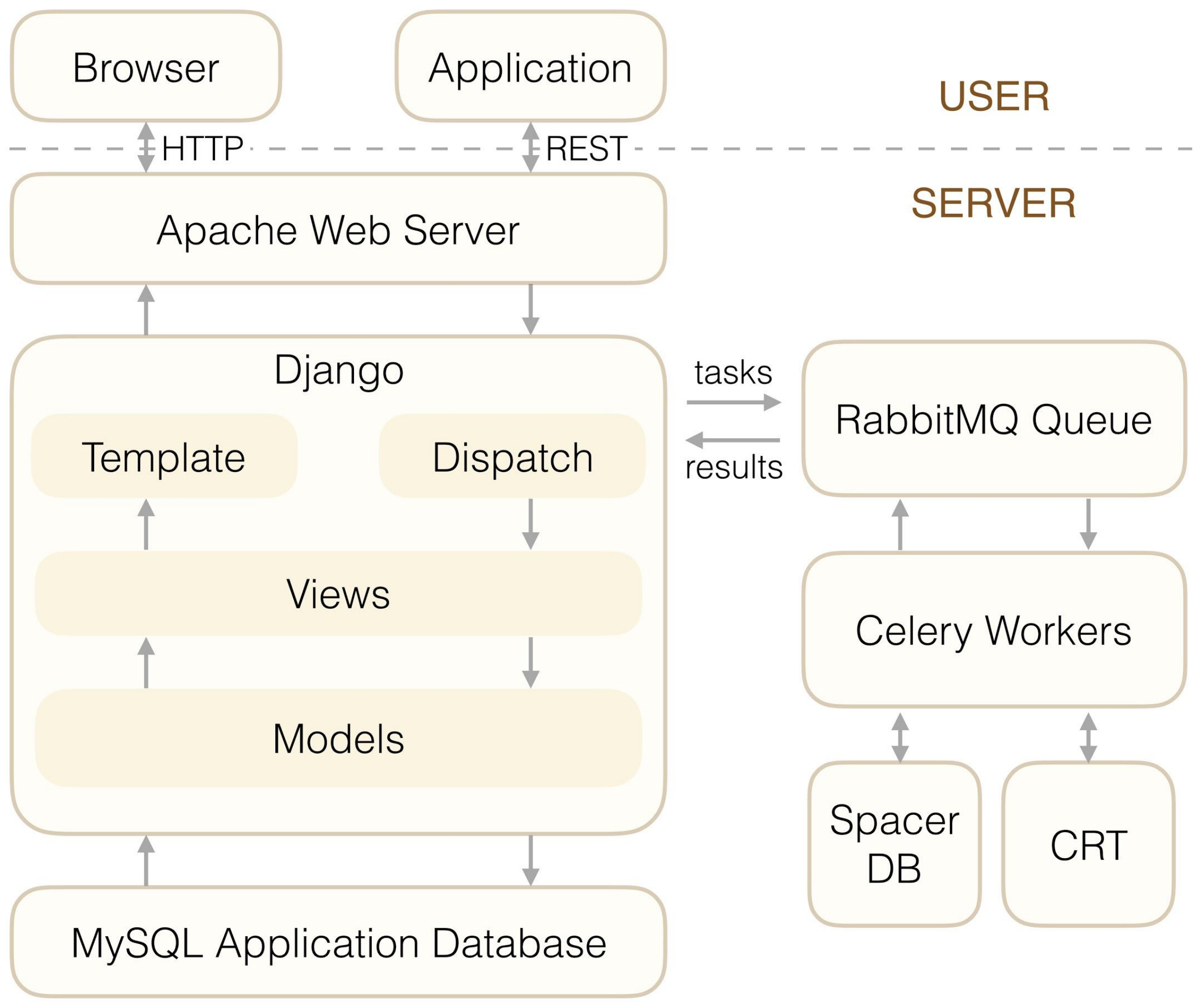
Finally, these CRISPR elements and the metadata fetched previously from NCBI genome database were used to create a relational database using MySQL.



Software architecture

The approach taken to expose the database and let the user interact with the data was to develop a web application. We used free and open source softwares to achieve that goal. We chose to publish our code on github, under a public license. Attention and care were given to reproducibility. As such, the environment was fully documented using an "infrastructure as code" approach.

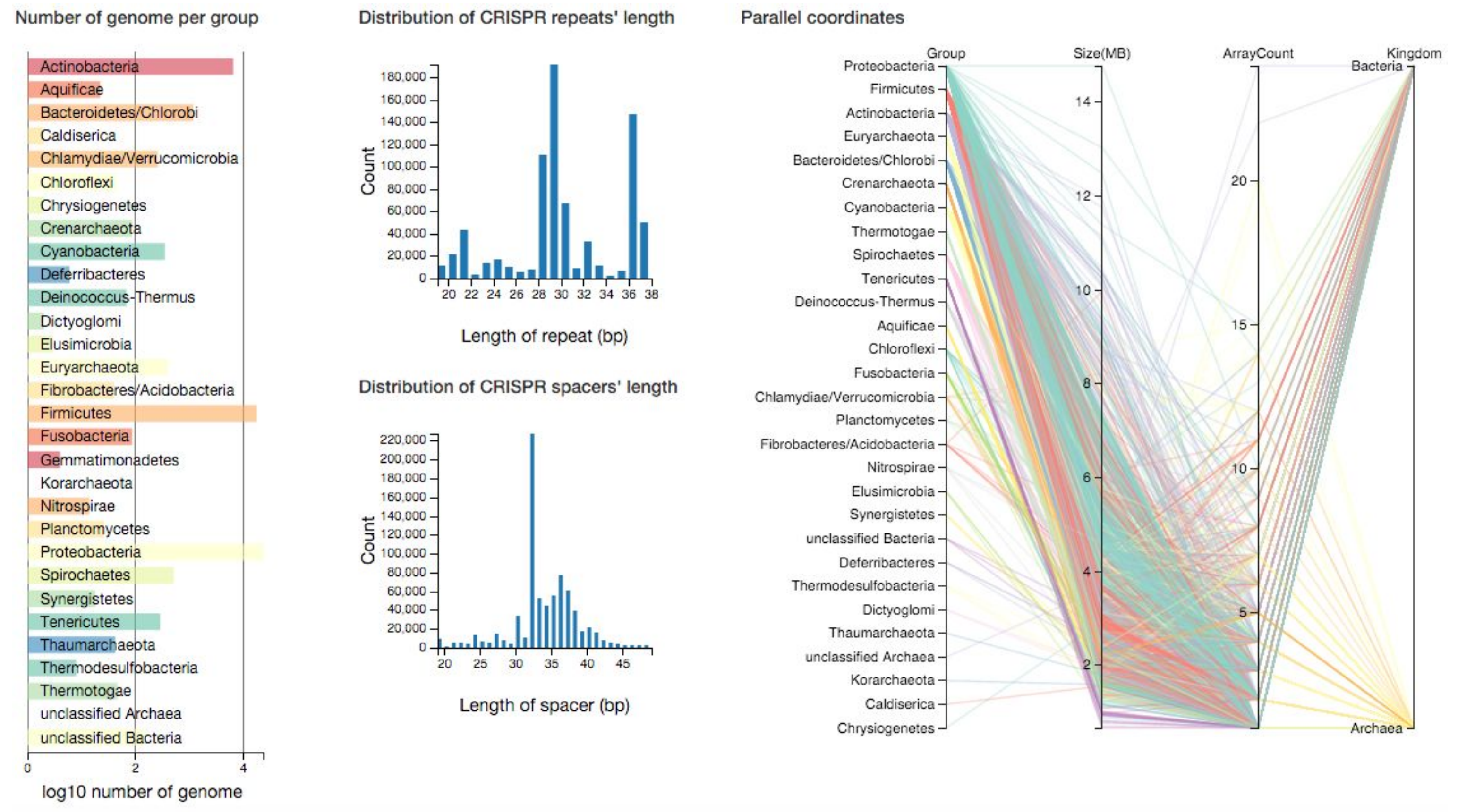
Repository address: https://github.com/Milt0n/CRISPR-Exposed
Website address: http://www.crispex.net/



Web application features

Users can interact with our application in multiple ways. Interactive data visualization allows for quick drill down into details of the dataset. Users can also programmatically interact with the database through a REST API. As services, we also provided an interface to blast sequences against our spacers database, and detect putative CRISPR array in their own sequences.

Data visualization



Discussion

The application and the database attached to it is novel in a couple of ways. To our knowledge, it is the first time such an extensive identification of CRISPR elements is conducted. About 54,000 genomes were used in this analysis. Other projects such as *CRISPRdb* [3], and *CRISPI* [4] only focus on the subset of genomes that have complete level assembly (4,000 genomes for *CRISPRdb*, 1,200 for *CRISPI*). As such, the spacer database constructed is the largest presently available online, consisting of 800,000 spacer sequences. We hope it can be useful to researchers, and can envision a couple of uses of our dataset:

- Comparative analysis of CRISPR arrays at species, class, or phylum level.
- Distribution of the origin of spacers in given bacteriophages, or plasmids.
- Phage host prediction.

Future work could include CRISPR-Cas systems identification and classification, PAM sequences information, as well as detection of anti-CRISPR genes. Time constraints prevented the integration of our early attempt at classification.

Authors' Contributions

Cedric, Hamed : Software architecture, developers, data processing
Hakim : Data visualization, literature studies
Ming : Data processing and visualization, CRISPR-Cas classification

[1] Barrangou, Rodolphe, et al. *Science* 315.5819 (2007): 1709-1712.
[2] Bland, Charles, et al. *BMC bioinformatics* 8.1 (2007): 209.
[3] Grissa, Ibtissem et al. *BMC Bioinformatics*. 2007 May 23;8(1):172
[4] Rousseau, Christine et al. *Bioinformatics*. 2009 Dec 15; 25(24): 3317–3318.
Complete references can be found on the dedicated page of the github repository: https://github.com/Milt0n/CRISPR-Exposed/blob/master/REFERENCES.md