

# A framework for mining GenBank: Implementation for the varDB project

Diego Diez<sup>1</sup>

diez@kuirc.kyoto-u.ac.jp

Nelson Hayes<sup>1</sup>

nelson@kuirc.kyoto-u.ac.jp

Nicolas Joannin<sup>2</sup>

nicolas.joannin@ki.se

Minoru Kanehisa<sup>1</sup>

kanehisa@kuirc.kyoto-u.ac.jp

Mats Wahlgren<sup>2</sup>

mats.wahlgren@ki.se

Craig E. Wheelock<sup>2</sup>

cewheelock@gmail.com

Susumu Goto<sup>1</sup>

goto@kuirc.kyoto-u.ac.jp

<sup>1</sup> Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011 Japan

<sup>2</sup> Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Box 280, SE-17177 Stockholm, Sweden

## Abstract

Sequence retrieval from repositories like GenBank is a difficult task when both comprehensiveness and accuracy are important. Sequence similarity methods may be used, at the expense of completeness. On the other hand, maintaining a database of papers publishing results on those sequences is also unsatisfactory because it is difficult to know all the existing publications for one gene family and prevents to obtain the huge amount of sequences deposited without related publication. Here, we present a framework that combines both approaches and apply it to an antigenic variant gene family.

## Keywords:

data mining, antigenic variation, gene families

## 1 Introduction

Data mining from sequence repositories is an open problem in Bioinformatics. For example, GenBank contains huge amounts of sequences that are submitted upon publication in scientific journals. One approach to have access to all sequences belonging to a gene family is to know the accession numbers of the papers where they were published and then get the sequences linked to it. However some publications contain data of several different gene families. Moreover, there are many sequences submitted without publication. Sequence similarity methods could be used, but they are based on arbitrary cutoffs that may be detrimental when both comprehensiveness and accuracy are the major goals.

The varDB [1] project aims to be a repository for antigenic variant sequences, covering different taxonomic groups and allowing cross-species studies. Related sequences come both from genome projects and from field isolate studies. Therefore, one of the goals of the varDB project is to provide an accurate and comprehensive description of the antigenic variant sequence space in GenBank. Here we present methods developed to solve this problem in the context of the varDB project. However, the methodology is generic and may be used for any other gene family of interest.

## 2 Materials and methods

The framework is developed in Perl and makes use of BioPerl and the NCBI EUtils. Sequences for specific taxa are downloaded in GenBank format from the different repositories, like the Nucleotide Core or ESTdb. Sequences belonging to genome projects are filtered out. Search for specific gene

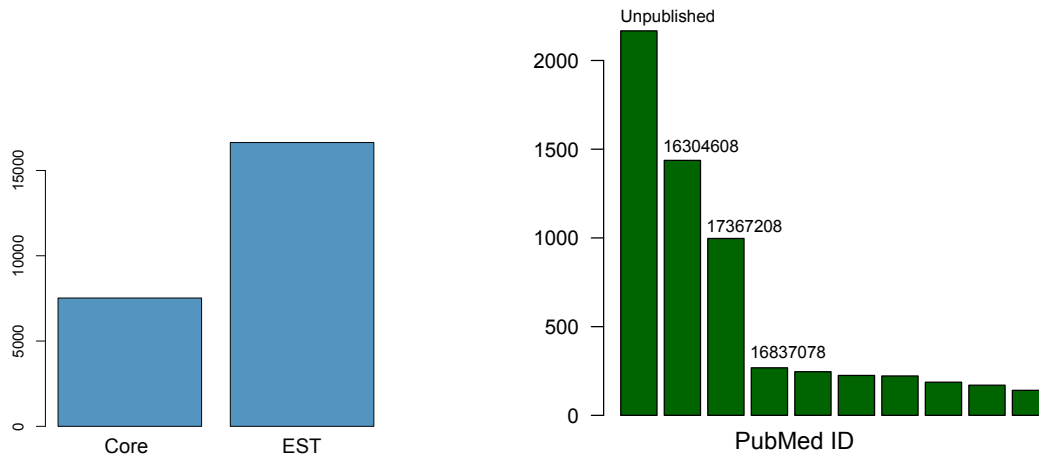


Figure 1: *var* sequences detected in the Core and EST branches of GenBank Nucleotide.

Figure 2: Partial *var* ESTs' sequence distribution among papers.

family sequences is performed with PSI-Blast, a profile similarity method. In parallel, a list of known publications related to the gene families of interest is maintained, which allows tuning of the search parameters and to included sequences not detected in the search step. As sequences associated with new publications are detected, these could be eventually reviewed and included in the publications list.

### 3 Results and Discussion

To illustrate the utility of the framework we present the results applied to the *var* gene family. This family encodes for a protein involved in antigenic variation in the Malaria parasite *Plasmodium falciparum*, and thousands of sequences have been deposited in GenBank. The initial publication list was extracted from Bull *et al.* The number of sequences detected in the Nucleotide Core and ESTdb repositories is showed in Figure 1. We wanted to see the coverage level, i.e. how many of the sequences belonging to known papers were successfully recovered. We found that on average, the coverage was higher than 90% indicating that sequence similarity methods cannot find all the sequences. In addition, we studied the publication distribution, showing in Figure 2 the results for ESTdb. Surprisingly, many detected sequences are not published and therefore the only way to find them is using sequence similarity methods. These results reveal the importance of using similarity methods combined with a database of known publications, in order to appreciate the whole set of sequences for a specific gene family.

### References

- [1] Hayes, C.N. *et al.*, varDB: a pathogen-specific sequence database of protein families involved in antigenic variation, *Bioinformatics*, (*in press*), 2008.
- [2] Bull, P.C. *et al.*, Plasmodium falciparum antigenic variation. Mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks, *Mol. Microbiol.*, 68(6):1519–34, 2008.