Clustering approach for identifying clonally-related sequences in immune repertoires

Clonal expansion, somatic hypermutation, and selection of B cell clones allow the immune system to produce millions of different types of Immunoglobulin (antibodies), highly specific in the defense against pathogens. High-throughput sequencing of B-cell immunoglobulin repertoires has provided large-scale data and opened new opportunities for studying the adaptive immune response in healthy individuals and those with a wide range of diseases. Identifying clones in B cell populations is central to detect the dysregulation that occurs with B-cell malignancies. It is also the starting point to several repertoire studies such as statistical analysis, repertoire comparisons, and clonal tracking. Several clustering methods have been developed to identify sets of clonally-related B-cells from high-throughput sequencing data. Most of them require the identification of VDJ genes involved in B-cell rearrangement, before undertaking clonal grouping. However, VDJ annotation is an arduous process, since millions of sequences need to be annotated, previously. We have proposed a simple approach to detect clonally related sequences without needing any gene annotation. Our method, named FaIR, clusters immunoglobulin sequences through their CDR3 region, the most variable part of immunoglobulins, used as a signature to detect faster sets of clonally related sequences. Our results shown that FaIR performs significantly better than most of the state of the art methods, and comparable to others. To identify clones FaIR uses greedy clustering algorithm, a extremely fast and still accurate clustering method. The objective of this project is to try more sophisticate clustering algorithms to identify clones such as those based on community detection [1, 2], or spectral clustering [3]. The student should implement or adapt such algorithms to identify clonally-related sequences in immune repertoires. We hope to improve the performance of FaIR without increase its computation time, since it should be compatible with clinical context.

References

1 - Blondel, V., et al, Fast Unfolding of Communities in Large Networks, Statistical Mechanics: Theory + Experiment, No 10, 2008

2 - https://github.com/taynaud/python-louvain

3 - C. Alpert, A. Kahng, and S. Yao. Spectral partitioning: The more eigenvectors, the better. Discrete Applied Math, 90:3- 26, 1999.