Yclon: Ultrafast clustering of B cell clones from high-throughput immunoglobulin repertoire sequencing data

Liza Figueiredo Felicori Vilela, Alice Ferreira de Souza, JOÃO HENRIQUE DINIZ BRANDÃO GERVÁSIO

UNIVERSIDADE FEDERAL DE MINAS GERAIS, UNIVERSIDADE FEDERAL DE MINAS GERAIS

Abstract

Next-generation sequencing technologies is revolutionizing our understanding about immunoglobulin (Ig) profile in different immune states. Clonotyping (grouping Ig sequences into B cell clones) is a way of investigating the diversity of repertoires, and how they change upon antigen exposure. Despite its importance, there is no consensus on the best method for clonotyping and the methods developed for that is computationally intractable for large sequencing datasets. This is the case of Change-O, that uses hierarchical clustering for this task. Because of this, we propose here to implement an approach to identify B cell clones from Ig repertoire data, named Yclon, focusing on reducing the runtime and computer memory usage. To do that Ig sequences sharing the same V and J gene segments and the same CDRH3 length were grouped, transformed in n-grams and then into a vector, weighed with tf-idf metrics and compared pairwise using cosine similarity. The resulted square distance matrix was the input for the Hierarchical DBSCAN (HDBSCAN) or for an alternative hierarchical agglomerative method. To test these 2 different methods developed here we used 3 Ig repertoire datasets (dataset 1 contains 82.927 sequences, the 2, 365.370 sequences and the 3, 1.741.413 sequences) and compared the results with the ones obtained by Change-O. For the 15% most abundant clones, regarding dataset 1 our results with HDBSCAN showed that 74% of the sequences within these clones were shared with Change-O, meanwhile using the agglomerative approach we observed 96.7% of shared sequences. This process took 13 seconds (HDBSCAN) and 14 seconds (agglomerative), while Change-O took 912 seconds in a computer with 8Gb RAM, Intel i5 core. For the dataset 2, our result with HDBSCAN shared 92% of the sequences with Change-O results, while with agglomerative 87% of the sequences were shared. For this dataset, Change-O took 4.280 seconds and Yclon made it in only 112 seconds (HDBSCAN) and 94 seconds (agglomerative). Importantly, Yclon were able to process a repertoire with 1.700.000 sequences in 6.300 seconds (HDBSCAN) and 2.422 seconds (agglomerative), but Change-O couldn't process this large dataset. Overall, we find that 2 different clustering approaches developed here, grouped Ig sequences into B cell clones as similar as Change-O did. However, we observed that Yclon was around 70 times faster and even was able to process larger than 1 million sequences Ig repertoire which is a critical part of repertoire studies and enables understanding the structure and affinity maturation of the repertoire.

Funding:
Link to Video: