BIO_21 - Construction and standardization of a self-updatable graph antibody database for the training neural networks to optimization of potential antiviral immunobiologicals

Diego da Silva de Almeida[1]; Geraldo Rodrigues Sartori[1]; João Hermínio Martins da Silva[1].
[1]Fundação Oswaldo Cruz (Fiocruz).

Introduction: Antibodies are immune system proteins produced by B lymphocytes that have been glimpsed as potent therapeutic candidates, being attractive for improved therapies. Its success comes from some characteristics of these molecules, such as low toxicity to the human organism and high affinity with their molecular targets. To evaluate the effectiveness of these compounds, a structural model is needed, and can be obtained by several methods. Computational techniques have advantages because they are cheaper and faster. This methodology has gained support from applications arising from artificial intelligence (AI) tools. Due to the pandemic scenario recently experienced, we started the development of a computational platform through deep learning (DL) that aims to identification and optimization of biopharmaceuticals, initially against COVID-19, being expandable for the confrontation with other diseases. For this purpose, the first step was the construction of a database to be used as a training model for the AI algorithm.

Objective: Based on that, the aim of this study is to implement an *in house* structural descriptor database of antibodies based on their biochemical and physicochemical properties, for use in graph-based DL routines for a biopharmaceutical optimization platform.

Methodology: Three-dimensional structures of antibodies' variable region complexed with their respective antigens were retrieved from AbDb database and standardized in sequence length and numbering. The structures were further converted to graph representation. For this, the characterization of the nodes and edges, fundamental units that form the graphs, occurred from specific molecular descriptors, such as the accessible surface area and weight binarization of distance between Cα for node and edge, respectively. One-hot encoding was used for categorical variable such as amino acid name and secondary structure.

Results: A total of 13 structural descriptors have been implemented to date.

Conclusion: More robust descriptors are being computed, and this database will be used for development of deep learning models.

Keywords: Artificial intelligence; Covid-19; Deep learning