Comparison Between a Graph-Based Metodology and the LSQKAB to Check Proteins Similarities

Monteiro, O.M., Dias, S.R., Rodrigues, T.S.

Centro Federal de Educação Tecnológica de Minas Gerais

Proteins are macromolecules present in all living beings and perform important functions, such as maintenance of organs and tissues, cell differentiation, transportation, and other several tasks. Many proteins have their three-dimensional structure solved and stored in biological databases, such as Protein Data Bank (PDB). There are softwares working with informations extracted from PDBs, some of these tools, has the function of checking similarities between protein structures. An example is LSQKAB, which belongs to CCP4 package and has its operation based on Kabsch algorithm, a technique frequently used in bioinformatics. Among the outputs of this software, is possible to get a list of the distances between each atoms pairs compared. However, this algorithm performs various computational calculations and it's use is inefficient to compare one record against a dataset. The main goal of this work is develop a methodology based on graphs to represent a protein structure in order to verify the similarity more efficient than LSQKAB, maintaining the accuracy. The first experiments were performed with 16,383 disulfide bounds files extracted from PDB. Each record contains 12 atoms and their distance values are represented by coordinates "x", "y" and "z". In the proposed methodology, each file is represented by a vector of 144 positions. Each position is refers to the calculation of Euclidean distance between the atomic distances of the 12 atoms. Using the representation, the clustering K-means technique was applied. The accuracy of each group generated was verified by LSQKAB trough of comparisons between all records of the same group. It was found that the more compact clusters correspond a more accurate result from LSQKAB, on the other hand, more spread clusters correspond a less accurate result from LSQKAB. This indicates that the behavior of the technique is similar to LSQKAB, moreover, more efficient on time consuming. The project is supported by Capes, CNPq and Fapemig.